

1 **EsoDetect: Computational Validation and Algorithm**
2 **Development of a Novel Diagnostic and Prognostic Tool for**
3 **Dysplasia in Barrett’s Esophagus**

4
5 Migla Miskinyte¹, Benilde Pondeca¹, José B. Pereira-Leal¹, Joana Cardoso¹

6 ¹ Ophiomics, Lisbon, Portugal

7

8 Corresponding Author:

9 Joana Cardoso

10 R. António Champalimaud 1 Sala 14, 1600-514, Lisbon, Portugal

11 Email address: jvaz@ophiomics.com

12

13

14

15

16

17 **ABSTRACT**

18 Barrett's esophagus (BE) is the only known precursor to esophageal adenocarcinoma (EAC), a
19 malignancy with increasing incidence and unfavorable prognosis. This study endeavors to
20 identify BE biomarkers capable of diagnosing low-grade dysplasia (LGD) in BE, as well as
21 biomarkers that can predict the progression from BE to EAC to be subsequently integrated into
22 diagnostic and prognostic algorithms.

23 Datasets containing gene expression data from metaplastic and dysplastic BE, as well as EAC
24 tissue samples, were collected from public databases and used to explore gene expression
25 patterns that differentiate between non-dysplastic (ND) and LGD BE (for diagnostic purposes)
26 and between non-progressed and progressed BE (for prognostic purposes). Specifically, for the
27 diagnostic application, three RNAseq datasets were employed, while for the prognostic
28 application, nine microarray datasets were identified, and 25 previously described genes were
29 validated. A Thresholding Function was applied to each gene to determine the optimal gene
30 expression threshold for group differentiation. All analyzed genes were ranked based on the
31 F1-score metrics. Following the identification of genes with superior performance, different
32 classifiers were trained. Subsequently, the best algorithms for diagnostic and prognostic
33 applications were selected.

34 In evaluating the value of gene expression for diagnosis and prognosis, the analyzed datasets
35 allowed for the ranking of biomarkers, resulting in eighteen diagnostic genes and fifteen
36 prognostic genes that were used for further algorithm development. Ultimately, a linear
37 support vector machine algorithm incorporating ten genes was identified for diagnostic
38 application, while a radial basis function support vector machine algorithm, also utilizing ten
39 genes, was selected for prognostic prediction. Notably, both classifiers achieved recall and
40 specificity scores exceeding 0.90.

41 The identified algorithms, along with their associated biomarkers, hold significant potential to
42 aid in the early management of malignant progression of BE. Their strengths lie in their
43 development using multiple independent datasets and their ability to demonstrate recall and
44 specificity levels superior to those reported in the existing literature. Ongoing experimental
45 and clinical validation is essential to further substantiate their utility and effectiveness, and to
46 ensure that these tools can be reliably integrated into clinical practice to improve patient
47 outcomes.

48

49 **KEYWORDS:** Barrett's esophagus; Esophageal adenocarcinoma; Biomarkers, Diagnosis;
50 Prognosis; Machine Learning

51

52 1. INTRODUCTION

53 Barrett's esophagus (BE) is characterized by the replacement of the normal squamous
54 epithelium lining the lower esophagus with specialized columnar cells (intestinal metaplasia)
55 [1–4]. This transformation occurs because of chronic gastroesophageal reflux disease (GERD)
56 [1, 2] and exposure to stomach acid [3]. Approximately 10% of patients with GERD are likely to
57 progress to a diagnosis of BE over 5 years [5]. Individuals with BE have a significantly increased
58 risk of developing esophageal adenocarcinoma (EAC). Typically, the progression of EAC starts
59 with GERD, followed by abnormal columnar cells characteristic of BE, which, over time, can
60 progress to dysplasia and eventually become EAC. Despite BE's role as a precursor to EAC, the
61 exact risk factors associated with BE are still not fully understood but include age (\geq 60-70
62 years), male gender [6], tobacco use [7, 8], obesity [9], and hiatal hernia [10].

63 The clinical relevance of BE relies on its role as the sole known precursor lesion for EAC [1, 11].
64 This specific type of esophageal cancer constitutes already around two-thirds of all cases of
65 esophageal cancer in high-income countries [12], with 85,700 new EAC cases estimated
66 worldwide in 2020. Over the next two decades, a staggering 65% increase (equivalent to
67 approximately 55,600 additional cases annually) is predicted [13]. EAC is a major problem
68 because of its association with poor survival rates, one of the lowest in oncology. Post-
69 diagnosis, EAC presents a 23% 5-year survival and a median survival of only 15 months [14],
70 highlighting the need for efficient methods for EAC management. This low survival is mainly
71 due to late diagnosis, limited treatment options, poor prognosis, high rate of early metastasis,
72 and difficulties in early detection [15].

73 Due to the low progression rate of BE to EAC (estimates 0.1-0.5, reviewed by [16]), most BE
74 patients never progress to cancer. However, GERD is becoming increasingly prevalent, with a
75 global estimate of 783 million prevalent cases in 2019 [17]. Factors like population growth,

76 aging, lifestyle changes, and improved living standards contribute to the rising incidence of
77 GERD [18]. As BE is a complication of GERD and a significant risk factor for EAC, the increasing
78 prevalence of GERD cases represents a menace to future management of EAC. The increased
79 prevalence of GERD leads to a higher incidence of BE cases and pressures for BE screening and
80 diagnosis, resulting in a significant economic burden for patients, families, health services, and
81 society.

82 Currently, BE serves as a critical warning sign and its surveillance is essential for effective risk
83 stratification. BE screening and surveillance methods involve endoscopic sampling of biopsies
84 from four quadrants according to the Seattle biopsy protocol [19, 20] followed by histological
85 analysis to classify detectable BE lesions as non-dysplastic (NDBE), indefinite for dysplasia
86 (IND), low-grade dysplasia (LGD), or high-grade dysplasia (HGD) [11]. Limitations to the success
87 of current strategies include but are not limited to, difficulties with endoscopic identification
88 of dysplasia, biopsy sampling error, low interobserver reproducibility in histologic assessment
89 of dysplasia among pathologists, lack of reliable biomarkers, access to specialized care and
90 patient compliance [21]. Variability in the endoscopic and histologic assessment are commonly
91 known issues: BE endoscopic/pathological management is time-consuming and depends on
92 the clinical experience of the physicians involved in the endoscopic examination and/or
93 histological analysis – who are mostly available in BE reference centers. For example, one meta-
94 analysis reported up to 25% and 24% of EACs were respectively missed during surveillance or
95 when the analysis was restricted to NDBE patients [22]. Regarding histological analysis, the
96 inter-observer agreement among pathologists has been reported as only 58% when it comes
97 to distinguishing normal esophagus from BE and was even lower (less than 50%) when
98 diagnosing LGD in BE patients [23, 24]. The lack of agreement can become particularly

99 problematic when many cases of BE are classified as IND (60% of dysplastic cases in a study by
100 Alshelleh et al. [25]) and when the interobserver agreement is even poorer than for LGD [26].
101 There is emerging evidence that the addition of biomarkers to risk stratification models could
102 increase BE diagnostic accuracy compared to current surveillance methods [27]. These
103 biomarkers range from the incorporation of more clinical variables [28, 29] to molecular
104 features such as genomic instability [30–34], gene expression patterns [35, 36], epigenetics
105 [37, 38], and proteomics [39]. In addition to biomarkers, the recent emergence of artificial
106 intelligence (AI) tools opens the prospect of improving the effectiveness of BE diagnosis and
107 surveillance. A recent meta-analysis revealed that deep learning algorithms applied to
108 endoscopy images in the surveillance of BE-related neoplasia are highly accurate (pooled
109 sensitivity and specificity of 90.3% and 84.4%, respectively) in detecting early HGD/EAC [40],
110 despite the absence of data for LGD. However, most diagnostic and prognostic tools
111 (biomarkers, AI), still lack substantial validation in large patient cohorts, refraining from their
112 usage in clinical practice [41]. In addition, the new tools available do not reach yet maximum
113 performance. For example, when predicting the neoplastic progression to HGD/EAC, both TP53
114 staining and Tissue Cypher test demonstrate high specificity (86% and 82%, respectively) but
115 to the detriment of low sensitivity/recall (49% and 55%, respectively) [reviewed by [42]].
116 While it is not yet clear whether regular surveillance surely leads to earlier detection of
117 dysplasia and consequently to a decrease in mortality from EAC [43] surveillance is still the only
118 recommended strategy for BE and EAC management. There is room for new diagnostic and
119 prognostic tools to support clinicians when diagnosing BE dysplasia and segmenting patients
120 based on the risk of BE progression to EAC.
121 The current study explores the diagnostic and prognostic value of gene expression patterns
122 from BE tissue samples from public datasets in the context of BE. Envisioning its clinical

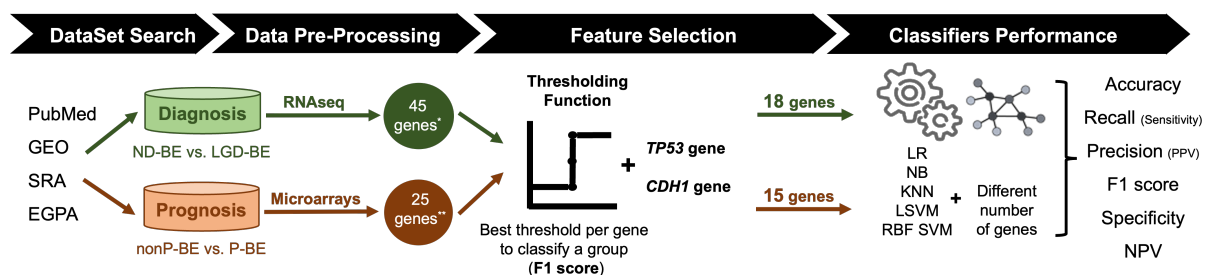
123 applicability, it aims to identify biomarkers that can accurately identify dysplasia within BE
124 lesions (diagnostic application) and biomarkers that can predict the progression to EAC
125 (prognostic application). It is also intended to understand the individual and combined
126 predictive value of each selected biomarker in both contexts through their implementation
127 using machine learning algorithms.
128

129 2. MATERIALS AND METHODS

130 2.1. Dataset Search:

131 An exhaustive search for public datasets containing gene expression data related to BE,
 132 including normal esophageal epithelium, NDBE, BE with different degrees of dysplasia (LGD
 133 and HGD) and EAC was performed in the following databases: Pubmed
 134 (<https://pubmed.ncbi.nlm.nih.gov/>), Gene Expression Omnibus (GEO,
 135 <https://www.ncbi.nlm.nih.gov/geo/>), Sequence Read Archive (SRA,
 136 <https://www.ncbi.nlm.nih.gov/sra/>), and European Genome-Phenome Archive ([https://ega-
 137 archive.org/](https://ega-archive.org/)). For the diagnostic application, the aim was to distinguish between NDBE and
 138 LGD BE. For the prognostic application, non-progressed BE (nonP-BE) and progressed BE (P-
 139 BE) data was studied. P-BE was defined as a BE adjacent to EAC. A summary of the methodology
 140 used is represented in **Figure 1** and described in detail below.

141



142

143 **Figure 1 – Methodology Summary.** Datasets of interest were downloaded from public databases, such
 144 as PubMed, GEO (Gene Expression Omnibus), SRA (Sequence Read Archive), and EGPA (European
 145 Genome-Phenome). For the diagnostic application, i.e., the distinction between non-dysplastic (ND) BE
 146 and low-grade dysplasia (LGD) BE, RNAseq datasets were used. Low-expression genes were excluded,
 147 resulting in a pre-selection of 45 genes. For the prognostic application, i.e., the distinction between
 148 non-progressed Barrett’s Esophagus (nonP-BE) and progressed-BE (P-BE), microarray datasets were
 149 identified, and 25 previously described genes were selected [35]**. A Thresholding Function was
 150 applied to each gene to define the best gene expression threshold for group distinction. All analyzed
 151 genes were ranked by F1 score, and additional feature selection methods were applied for diagnostic
 152 genes*, determining the top genes for diagnosis and prognostic application. Due to their biological
 153 functions, two extra genes – *TP53* and *CDH1* – were added to both diagnostic and prognostic data sets,
 154 summing 18 diagnostic and 15 prognostic genes. Different algorithms – Logistic Regression (LR), Naive
 155 Bayes (NB), K-nearest neighbours (KNN), Linear Support Vector Machines (LSVM), and Radial Basis
 156 Function Support Vector Machines (RBF SVM) – were trained using different numbers of genes.

157 Algorithms' performance was assessed through accuracy, recall (sensitivity), precision (positive
158 predictive value – PPV), F1 score, Specificity and negative predictive value (NPV). [35]
159

160 2.2. Data Pre-Processing:

161 In this study, raw RNA-seq data from projects GSE193946, GSE58963, and E-MTAB-4054 were
162 obtained from the Sequence Read Archive (SRA) and the European Genome-Phenome Archive
163 (EGA). We processed the data using a Docker environment equipped with Kallisto version
164 0.46.1 (docker image: jlnetosci/kallisto:v0.46.1), which facilitated the pseudo-alignment of the
165 reads against the Homo_sapiens.GRCh38.cdna.all.release-107 reference transcriptome from
166 Ensembl. Post-alignment, the transcript abundance estimates generated by Kallisto were
167 imported into the R programming environment using the tximport package. This allowed
168 transcript-level data to be transformed into gene-level counts, which were subsequently
169 analyzed for differential expression. The combined data was filtered for low-expressed genes
170 using the filterByExpr function in EdgeR [44], resulting in a dataset of 20,608 genes for
171 downstream analysis. Samples were then normalized using the TMM (Trimmed Mean of M-
172 values) normalization method and differential expression analysis was performed using EdgeR
173 [31]. For downstream analysis, including feature selection and classifier training, log-
174 transformed CPM normalized values were used, which were subsequently corrected for batch
175 effects using the ComBat function from the sva package [45].

176 In this study, microarray data was sourced from the Gene Expression Omnibus (GEO) database
177 using the GEOquery package available in the R software. The data included accessions
178 GSE1420, GSE363223, GSE13083, GSE37200, GSE34619, GSE26886, GSE39491, GSE100843,
179 and an additional dataset from Watts et al. (2007) [46]. Data was loaded and normalized using
180 both the affy and oligo packages in R, depending on the array platform. The CEL files were read
181 and processed using the frma function for robust multi-array average (RMA) normalization.

182 Probe-level data was annotated and collapsed to gene-level data using Bioconductor
183 annotation packages `hugene10sttranscriptcluster.db`, `hgu133a.db`, `hgu133plus2.db`, and
184 `hgu133a2.db` along with the WGCNA package. Finally, the resulting gene expression data was
185 merged into a single dataset for downstream analysis, with additional annotations indicating
186 BE progression status [42][43]. For prognostic application, 25 genes selected in previous work
187 to distinguish nonP-BE from P-BE [35], were used in this study – *ACTN1*, *C1S*, *CCN1* (alias
188 *CYR61*), *CDH1*, *CEBPB*, *CEBPD*, *COL4A1*, *CTSB*, *DKK3*, *DUSP1*, *IER3*, *JUN*, *LAMC1*, *PLPP3*, *RBPM5*,
189 *SNAI1*, *SNAI2*, *SPARC*, *TNS1*, *TRMT112*, *TP53*, *TWIST1*, *VWF*, *WWTR1* (alias *TAZ*) and *ZEB1*. Box
190 plots representing normalized expression values were generated using the `ggplot2` (v3.4.0) and
191 `ggsignif` (v0.6.4) R packages. Statistical analysis was performed using one-way ANOVA, followed
192 by a *post hoc* Tukey's 'Honest Significant Difference' test, both from the R stats package
193 (v4.1.1). When ANOVA assumptions were not met, a Kruskal-Wallis Rank Sum Test (R stats
194 package v4.1.1) was performed, followed by a *post hoc* Dunn's Kruskal-Wallis Multiple
195 Comparisons test (FSA R package v0.9.3). The significance threshold was set at *p-value* < 0.05.

196

197 2.3. Threshold selection and determination of individual predictive power:

198 For the distinction between NDBE and LGD (diagnostic) or nonP-BE and P-BE (prognostic) a
199 Thresholding function was applied to the expression levels of each selected gene to determine
200 an expression threshold. Performance metrics such as accuracy, recall (or sensitivity), precision
201 (or positive predictive value – PPV), specificity, negative predictive value (NPV), and false
202 positive rate (FPR) were calculated for each threshold, considering the known class of the
203 samples. For the diagnostic application, other feature selection methods (Lasso, Mutual
204 Information (MI) criteria, Recursive Feature Elimination (RFE), SelectKBest) were also applied
205 to narrow down the most informative features that appeared at least twice in one of the

206 methods. The threshold that yielded the highest F1-score was selected. Based on this metric,
207 genes were ranked and the top 16 (diagnostic) and top 13 (prognostic) were considered for
208 downstream analysis. Two additional genes – *TP53* and *CDH1* – were also included in the
209 downstream analysis of both prognostic and diagnostic gene sets.

210

211 2.4. Algorithmic analysis and evaluation of performance metrics:

212 Gene expression values were used for algorithm training. Several classes of classifiers, with
213 shown applicability to microarray and RNAseq data [47–49], such as Logistic Regression (LR),
214 Naive Bayes (NB), K-nearest neighbours (KNN), and Support Vector Machines (SVM) (with
215 Linear and Radial Basis Function kernels), were implemented with default hyperparameters in
216 Python programming language (v3.10.0), using the scikit-learn package (v1.0.1). A leave-one-
217 out cross-validation procedure was used to evaluate the diagnostic or prognostic value of all
218 possible combinations of genes (from n=2 up to all selected diagnostic or prognostic genes).
219 This involved leaving out one sample at a time for validation while using the remaining samples
220 to create a balanced training set. The Synthetic Minority Oversampling Technique (SMOTE)
221 was employed from the imbalanced-learn (v0.8.1) package. For LR, KNN, and SVM, features
222 were standardized (scaled and centered) using scikit-learn's standard scaler module by
223 subtracting the mean and scaling to the unit variance. Performance metrics such as accuracy,
224 precision (PPV), recall (sensitivity), NPV, and precision and specificity were calculated and
225 recorded for each full iteration of the validation strategy. The top-performing algorithms were
226 chosen by maximizing performance metrics (accuracy, specificity, precision, recall, NPV, and
227 F1-score, Table 2). The most frequent models, with the highest F1-score, were chosen to
228 further select the best classifiers for both diagnostic and prognostic applications. The most
229 frequently occurring genes (frequency $\geq 50\%$) within the selected classifiers were chosen as

230 features. Subsequently, the performance metrics were calculated using a decremental number
231 of features, and the median value and standard deviation of each group of decremental
232 subsets of genes were computed.

233

234 2.5. *In-vivo* gene expression analysis:

235 2.5.1. RNA extraction:

236 Cell pellets from the cell lines metaplasia (BAR-T and BAR-T10 - from R. Souza, Baylor
237 University Medical Center, Dallas, TX; Jaiswal et al., 2007; X. Zhang et al., 2010), dysplasia (CP-
238 B, CP-C and CP-D - from P. Rabinovitch, University of Washington, Seattle, WA; Palanca-
239 Wessels et al., 2003), and EAC (OE33, KYAE-1- from W. Dinjens, Erasmus Medical Center
240 Cancer Institute, Rotterdam, Netherlands, and ESO26 - Boonstra et al., 2010) were used to
241 extract RNA using the RNeasy Mini Kit (#74104, Qiagen, Hilden, Germany), following the
242 manufacturer's instructions.

243 For formalin-fixed paraffin-embedded (FFPE) tissue samples, RNA was isolated from 2
244 consecutive sections per sample, each approximately 20 mm² and 5 µm. Tissue samples
245 were deparaffinized using the deparaffinization solution (#19093, Qiagen, Hilden, Germany)
246 prior to RNA extraction with the RNeasy FFPE Kit (#73504, Qiagen, Hilden, Germany),
247 according to the manufacturer's instructions (with one modification: proteinase K incubation
248 was performed overnight).

249 All procedures involving human tissue samples were approved by the National Ethics
250 Committee for Clinical Research – Comissão de Ética para a Investigação Clínica (CEIC), under
251 approval number 2022_EO_24.

252

253 2.5.2. Reverse Transcription - quantitative real-time Polymerase Chain Reaction (RT-
254 qPCR):

255 For 1-Step RT-qPCR, reactions were performed in triplicate, using the TaqPath 1-step RT-qPCR
256 Master Mix (#A15300, Thermo Fisher Scientific) with a final reaction volume of 10 μ L. Each
257 reaction containing 1 μ L of template, 0.25 μ M of probe and 0.5 μ M of each primer. Data
258 acquisition and analysis were conducted using the QuantStudio Design & Analysis Software
259 v1.5.1 software, using the cycling program: UNG incubation at 25°C - 2 minutes, Reverse
260 Transcription at 50°C - 15 minutes, followed by Polymerase activation at 95°C - 2 minutes and
261 40 cycles of Amplification at 95°C - 3 s and 58°C - 30 s. To normalize gene expression levels,
262 the geometric mean of the reference genes (*PGK1*, *ELF1*, and *RPL13A*) was subtracted from
263 cycle threshold (Cq) of the target genes.

264

265 **3. RESULTS:**

266 **3.1. Diagnosis and prognosis dataset selection**

267 For the development of the diagnostic application, 13 RNAseq-based datasets were identified,
268 of which only 3 had publicly available clinical data – GSE58963 [50], E-MTAB-4054 [51],
269 GSE193946 [52] – and were therefore included in the present study. The BE data contained in
270 each dataset is represented in **Table 1**. In total, data from 61 samples – comprising 21 NDBE,
271 40 LGD BE and 27 HGD – were included in the study.

272

273 **Table 1 – Characterization of datasets for the diagnostic and prognostic applications**

Dataset	Diagnostic (RNAseq)			Prognostic (Microarray)	
	NDBE	LGD	HGD	nonP-BE	P-BE*
GSE1420 [53]	-	-	-	0	16
Watts 2007 [46]	-	-	-	18	0
GSE36223 [54]	-	-	-	23	0
GSE13083 [55]	-	-	-	7	0
GSE37200 [56]	-	-	-	0	46
GSE34619 [57]	-	-	-	10	0
GSE26886 [58]	-	-	-	20	0
GSE39491 [59]	-	-	-	40	0
GSE100843 [60]	-	-	-	17	3
GSE58963 [50]	7	7	7	-	-
E_MTAB_4054 [51]	14	8	-	-	-
GSE193946 [52]	0	25	20	-	-
TOTAL N. samples	21	40	27	135	65

274 nonP-BE – non progressed Barrett’s esophagus, P-BE – progressed Barrett’s esophagus, NDBE – non-
275 dysplastic Barrett’s esophagus,

276 LGD – low-grade dysplasia, HGD – high-grade dysplasia.

277 *P-BE was defined when a BE was adjacent to EAC

278

279 For the prognostic application, 16 microarray datasets were identified, but only those
280 generated on an Affymetrix platform were included in the downstream analysis to facilitate
281 data merging. A total of 9 microarray datasets were analyzed, including three previously
282 analyzed by Cardoso, *et al* [35], – GSE1420 [53], Watts_2007 [46], and GSE13083 [55] and six
283 new ones, namely GSE36223 [54], GSE37200 [56], GSE34619 [57], GSE26886 [58], GSE39491

284 [59], and GSE100843 [60]. In total, data from 200 samples – representing 135 nonP-BE and 65
285 P-BE – were included in the study as shown in **Table 1**.

286

287 **3.2. Identification of differentially expressed genes in a diagnostic and** 288 **prognostic setting**

289 In this study, we aimed to identify diagnostic biomarkers that can distinguish between ND-BE
290 and LGD-BE. For this purpose, we utilized three RNAseq datasets (as listed in **Table 1**). Low-
291 expression genes were excluded from each dataset, resulting in the inclusion of 20608 genes
292 in our analysis. After normalization, we conducted differential expression analysis between
293 LGDBE and NDBE (Figure S1A), and HGDBE and NDBE (Figure S1B) using EdgeR's quasi-
294 likelihood approach. This approach accounted for disease staging and batch effects from the
295 three different datasets as factors in the model (Supplementary Table 1, Figure S1C).

296 Following the differential expression analysis, we identified 30 biomarkers through a
297 systematic selection process. First, we selected differentially expressed genes (DEGs) with an
298 absolute log fold change (logFC) of ≥ 1 between LGDBE and NDBE, with a false discovery rate
299 (FDR) of < 0.05 . From these DEGs, we filtered for genes that showed the same direction of
300 expression change in the HGDBE vs. NDBE comparison (FDR < 0.05), resulting in 14 genes
301 (Figure S1A). Second, we identified DEGs in the HGDBE vs. NDBE comparison with an absolute
302 logFC of ≥ 2 (FDR < 0.05). Among these genes, we selected those that also exhibited the same
303 direction of expression change in the LGDBE vs. NDBE comparison (considering p-value < 0.05
304 for significance), resulting in 16 genes (Figure S1B). This two-step filtering strategy ensured
305 that the selected biomarkers not only had significant differential expression but also consistent
306 expression patterns across different stages of disease progression. Given the strong batch
307 effect observed (see Figure S1), there was a risk of losing biologically relevant genes in the

308 LGDBE vs. NDBE comparison due to this variation. To mitigate this problem, we also performed
309 separate analyses of the EMTAB_4054 (Supplementary Table 2) and GSE58963
310 (Supplementary Table 3) datasets. We employed the glmRobust pipeline to independently
311 identify differentially expressed genes between the LGDBE and NDBE groups within each
312 dataset. From these separate analyses, we identified an additional 13 genes with an absolute
313 logFC greater than 1 and an FDR < 0.05. These genes were consistently found in both datasets
314 and exhibited the same direction of expression change (Figure S2). Moreover, these genes
315 showed consistent directional changes in the previous HGDBE vs. NDBE comparison. Thus, they
316 were also included in the biomarker list (Supplementary Table 4). Given their established role
317 in the biology of BE and EAC, we also included two additional genes – *TP53* and *CDH1* –in the
318 downstream analysis, resulting in a total of 45 candidate genes for distinguishing between
319 NDBE and LGD.

320 For the prognostic set of biomarkers, we re-analyzed 25 genes that we had previously
321 identified to have prognostic value [35], namely *ACTN1*, *C1S*, *CCN1* (alias *CYR61*), *CDH1*, *CEBPB*,
322 *CEBPD*, *COL4A1*, *CTSB*, *DKK3*, *DUSP1*, *IER3*, *JUN*, *LAMC1*, *PLPP3*, *RBPM5*, *SNAI1*, *SNAI2*, *SPARC*,
323 *TNS1*, *TP53*, *TRMT112*, *TWIST1*, *VWF*, *WWTR1* (alias *TAZ*) and *ZEB1*. For validation purposes,
324 we added six independent datasets to the three datasets we originally analyzed. We observed
325 significant differential gene expression (adj. *p-value* < 0.05) between P-BE and nonP-BE
326 categories for most of the genes of interest, except for *CDH1*, *DKK3*, *SNAI2*, and *WWTR1*.

327

328 3.3. Application of a Thresholding function for the selecting genes with the 329 highest predictive value

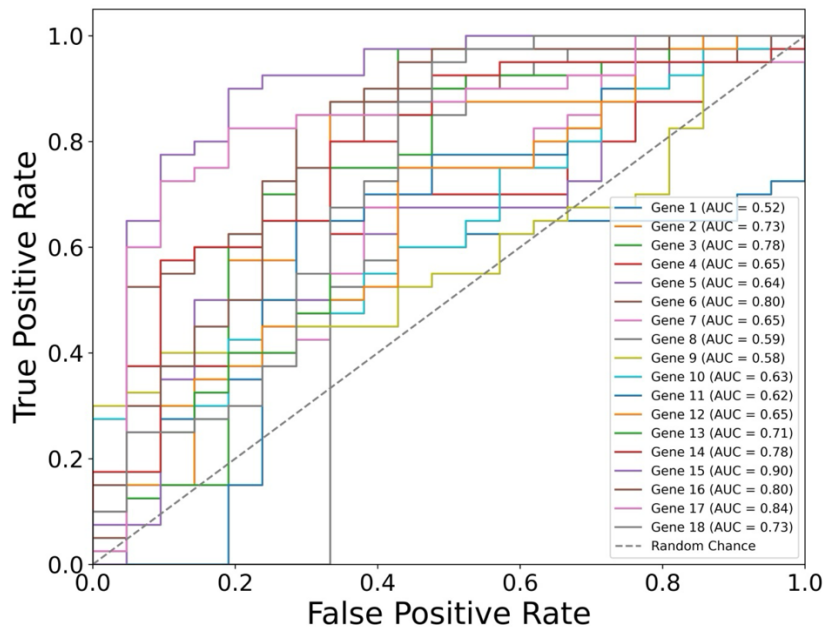
330 To each selected gene, we applied a Thresholding function, to determine a gene expression
331 threshold for distinguishing different levels of gene expression between groups of samples

332 with distinct diagnosis (NDBE vs. LGD-BE) or with distinct prognosis (nonP-BE vs. P-BE). We
333 defined the best individual threshold of gene expression for each selected gene, 45 for
334 diagnosis and 25 for prognosis, reflecting the individual predictive value of each gene. Genes
335 were then ranked by the harmonic mean of recall and precision (F1-score) to ensure accurate
336 selection. This procedure identified the top 15 genes for predicting the malignant progression
337 of BE lesions with F1-score above 0.67 (Supplementary Table 5). From the top 45 diagnostic
338 genes, including CDH1 and TP53, genes with higher expression values (log2CPM above 1) were
339 filtered. To further refine a list of candidates, we used several feature selection methods: Lasso,
340 Mutual Information (MI) criteria, Recursive Feature Elimination (RFE), and SelectKBest.
341 Additionally, feature correlation analysis was conducted to identify and eliminate highly
342 correlated features (Pearson's correlation coefficient > 0.9). Hence, for diagnostic purposes,
343 we further narrowed down the selection to genes that were chosen at least twice in one of the
344 feature selection methods and F1-score above 0.7, which identified the top 16 genes for
345 diagnosing dysplasia in the context of BE (Supplementary Table 6).

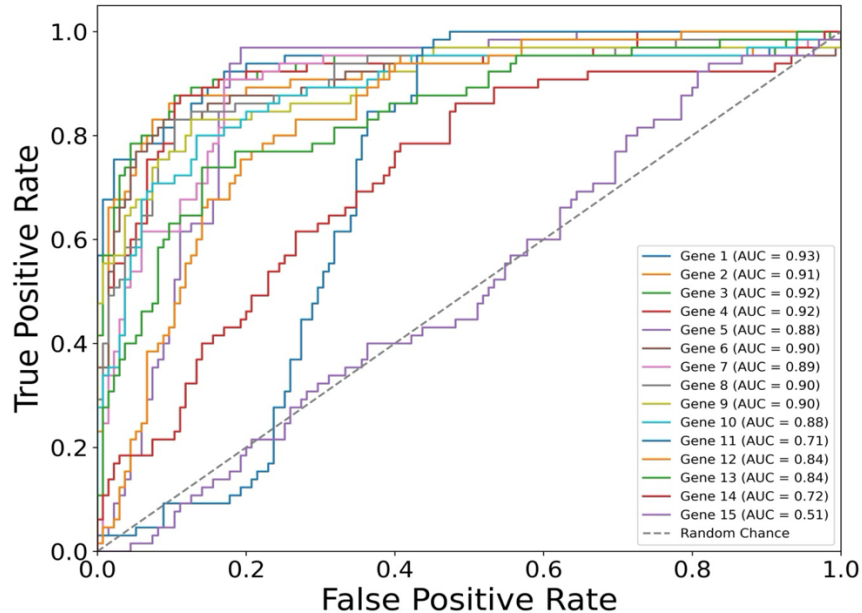
346 Building on our identification of the top diagnostic and top prognostic genes using the
347 Thresholding function and various feature selection methods, the ROC curve results (Figure 2)
348 further validate their predictive power using a logistic regression classifier. In both diagnostic
349 (Figure 2A) and prognostic (Figure 2B) contexts, most of the genes have AUC (Area Under the
350 Curve) values above 0.50 (random chance line), with some reaching individual values of 0.90,
351 demonstrating a substantial predictive value of the selected genes in both diagnostic and
352 prognostic applications.

353

A Diagnostic (dysplasia)



B Prognostic (progression)



354

355 **Figure 2 – Gene-Specific ROC Curves for Diagnostic and Prognostic Predictions.** Receiver Operating
356 Characteristic (ROC) curves for individual genes depicting their predictive value in two contexts: **(A)**
357 Diagnostic (dysplasia) and **(B)** Prognostic (progression) using a logistic regression classifier. The Area
358 Under the Curve (AUC) values for each gene are indicated in the legends. Notably, the predictive values
359 of TP53 and CDH1 genes are also included, although they were manually added to the sets.

360

361 **3.4.SVM algorithms were the best for both diagnostic and prognostic applications**

362 The diagnostic and prognostic gene groups were utilized to train the most effective diagnostic and
363 prognostic algorithms. Various classifiers – LR, NB, KNN, LSVM, and RBF SVM – were examined using
364 increasing combinations of genes, ranging from n = 2 up to the total number, for diagnostic and
365 prognostic applications.

366 The algorithms were ranked based on their performance metrics for each application (see **Table 2**).
367 However, no algorithms optimized all performance metrics for both applications. Nevertheless, the
368 LSVM algorithms emerged as the best for diagnostic purposes, maximizing the F1-score and accuracy
369 (refer to **Figure 3A** and **Table 2**).

370 For the prognostic application, a similar trend was observed, where the RBF SVM type performed best
371 according to the F1-score and accuracy metrics (refer to **Figure 3B** and **Table 2**).

372

373 **Table 2 – Best algorithm performance by metric maximization**

Application	Rank by	N. algorithms	Type of algorithm	Recall	Precision	F1 - score	Specificity	NPV	Accuracy
Diagnostic	Recall	4871	KNN (n=196) LSVM (n=2426) LR (n=124) RBF SVM (n=2125)	0.99	0.78-0.98	0.88-0.99	0.48-0.95	1.00	0.82-0.98
	Precision	3050	KNN (n=2290) LSVM (n=472) LR (n=259) NB (n=21) RBF SVM (n=8)	0.65-1.00	0.97	0.79-0.99	0.95-1.00	0.60-1.00	0.77-0.98
	F1-score	1881	KNN (n=288) LSVM (n=1115) LR (n=444) RBF SVM (n=34)	0.92-1.00	0.93-1.00	0.96	0.86-1.00	0.88-1.00	0.95-0.98

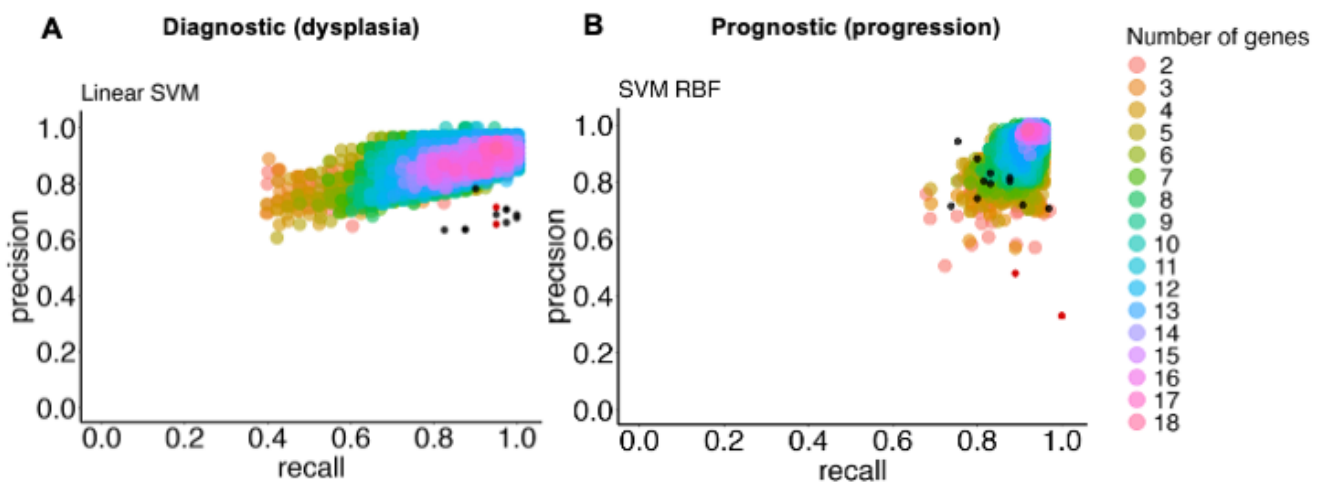
Application	Rank by	N. algorithms	Type of algorithm	Recall	Precision	F1 - score	Specificity	NPV	Accuracy
	Specificity	231	KNN (n=223) LSVM (n=8)	0.65-0.95	1.00	0.79-0.97	0.99	0.60-0.91	0.77-0.97
	NPV	4871	KNN (n=196) LSVM (n=2426) LR (n=124) RBF SVM (n=2125)	1.00	0.78-0.98	0.88-0.99	0.48-0.95	0.99	0.82-0.98
	Accuracy	212	KNN (n=38) LSVM (n=157) LR (n=16) RBF SVM (n=1)	0.95-1.00	0.95-1.00	0.97-0.99	0.90-1.00	0.91-1.00	0.96
Prognostic	Recall	13	LR (n=7), LSVM (n=5), RBF SVM (n=1)	0.97	0.69-0.70	0.81	0.79-0.80	0.98	0.85-0.86
	Precision	582	RBF SVM (n=449) KNN (n=24) LSVM (n=17) NB (n=92) LR (n=348)	0.88-0.95	0.99	0.93-0.98	1.00	0.94-0.98	0.96-0.98
	F1-score	12971	RBF SVM (n=5794) KNN (n=2465) LSVM (n=2230) NB (n=2134) LR (n=348)	0.92-0.95	0.97-1.00	0.96	0.99-1.00	0.96-0.98	0.98
	Specificity	8430	KNN (n=586) LSVM (n=569) LR (n=38)	0.83-0.95	0.98-1.00	0.9-0.98	0.99	0.92-0.98	0.94-0.98

Application	Rank by	N. algorithms	Type of algorithm	Recall	Precision	F1 - score	Specificity	NPV	Accuracy
			NB (n=1953) RBF SVM (n=5284)						
	NPV	13	LR (n=7), LSVM (n=5), RBF SVM(n=1)	0.97	0.69-0.70	0.81	0.79-0.80	0.98	0.85-0.86
	Accuracy	2404	KNN (n=264) LSVM (n=370) LR (n=28) NB (n=415) RBF SVM (n=1327)	0.94-0.95	0.98-1	0.97-0.98	0.99-1	0.98	0.98

374 LR – logistic regression, LSVM – linear support vector machine, RBF SVM – radial basis function support vector

375 machine, KNN – K-nearest neighbors, NB – Naïve Bayes. Selected algorithms are highlighted in grey.

376



377
378 **Figure 3 - Precision and Recall for the selected classifier type with increasing combinations of genes of**
379 **interest to predict BE dysplasia (diagnostic) and BE malignant progression (prognostic).** This illustrates
380 the performance of the chosen classifier types in the predicting dysplasia (A) and progression (B) when
381 different numbers of genes of interest are combined (colored dots). The individual predictive value for the
382 best threshold of each previously selected gene is also represented (black dots) for diagnostic (see **Figure**
383 **3A**) and prognostic (see **Figure 3B**). Colors represent different numbers of combined genes. LSVM – Linear
384 Support Vector Machine (A), and RBF SVM – Radial Basis Function Support Vector Machine (B). Red dots
385 represent manually added CDH1 and TP53 genes.

386

387 The study found that among the selected types of algorithms, those with an F1-score above
388 0.96 included 1115 LSVM for diagnostic and 5794 RBF SVM for prognostic. The analysis
389 identified the most frequent genes (over 50 %) across the best-performing algorithm class
390 (Supplementary Table 7 and Supplementary Table 8). Ultimately, ten genes were selected for
391 identifying LGD BE using a LSVM algorithm: *IGHV3-43*, *SLC38A4*, *PLLP*, *CELA3A*, *IGHV4-31*,
392 *TMPRSS5*, *TP53*, *NR4A1*, *ATF3*, *IFI27*. For identifying P-BE, ten genes were selected using an
393 RBF SVM algorithm: *SNAI1*, *C1S*, *DUSP1*, *CEBPB*, *COL4A1*, *ZEB1*, *CEBPD*, *CCN1*, *LAMC1* and
394 *TWIST1*.

395

396 The performance of each selected algorithm (LSVM for diagnosis and RBF SVM for prognosis)
397 was evaluated using the most frequent genes (10 for diagnosis and 10 for prognosis) as
398 features. To test different random states while avoiding algorithm bias, 100 runs were
399 performed for each algorithm with the same features. **Table 3** presents the mean values and
400 respective standard deviations (SD) for each performance metric. All performance metrics
401 were above 0.90, except for specificity for the LSVM diagnostic algorithm. The low standard
402 deviations (below 0.05) indicated an increase in the predictive value of each algorithm when
403 the selected genes were combined.

404

405

406

407

408

409

410

411 **Table 3 – Performance of the selected algorithms with the selected genes as features after 100 runs**

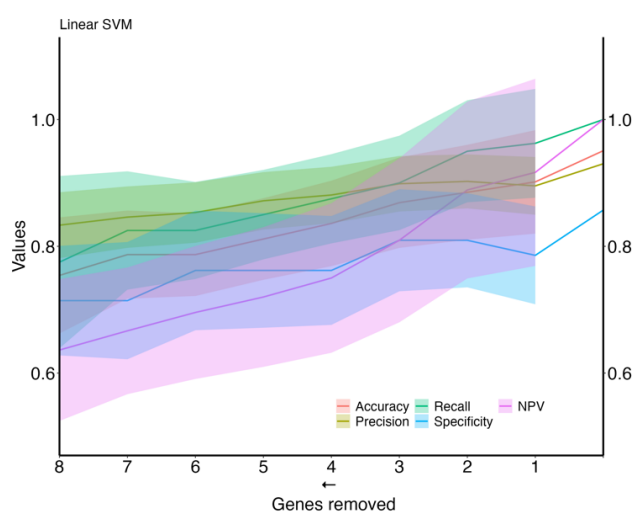
A. Diagnostic			B. Prognostic		
	Mean	Standard deviation		Mean	Standard deviation
Accuracy	0,946	0,014	Accuracy	0,977	0,003
Precision	0,932	0,012	Precision	0,977	0,008
Recall	0,991	0,017	Recall	0,952	0,005
F1 score	0,960	0,010	F1 score	0,965	0,004
TP	39,630	0,677	TP	61,900	0,302
FP	2,900	0,541	FP	1,430	0,498
TN	18,100	0,541	TN	133,570	0,498
FN	0,370	0,677	FN	3,100	0,302
NPV	0,981	0,033	NPV	0,977	0,002
Specificity	0,862	0,026	Specificity	0,989	0,004
FPR	0,138	0,026	FPR	0,011	0,004

412 TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative, NPV- Negative Predictive Value, FPR-
413 False Positive Rate.

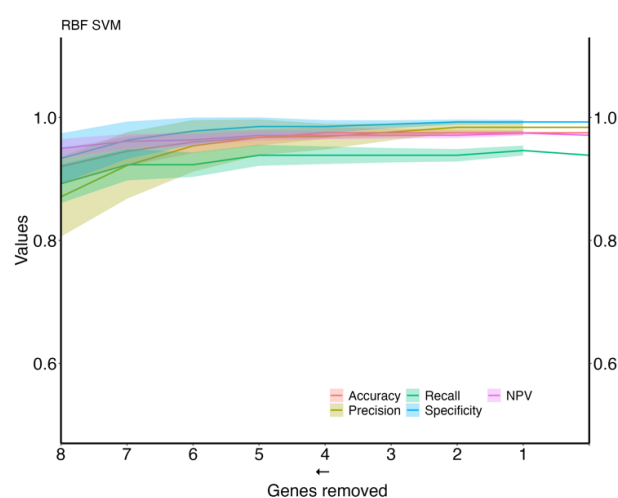
414
415 Finally, the performance of the two algorithms was evaluated by gradually decreasing the
416 number of selected genes (**Figure 4**). The diagnostic algorithm showed a decrease in
417 performance after the removal of just one gene (**Figure 4A**). In contrast, the prognostic
418 algorithm showed noticeable changes only after the removal of four genes (**Figure 4B**).

419

A Diagnostic (dysplasia)



B Prognostic (progression)



420 **Figure 4 – Metrics performance of the best algorithm in distinguishing ND-BE and LGD BE**
421 **(diagnostic) and nonP-BE from P-BE (prognostic) when decreasing the number of genes included in**

422 **the training.** Mean of each performance metric (solid lines) and its respective standard deviation
423 (ribbons) for the diagnostic algorithm, Linear Support Vector Machine (LSVM) **(A)** and for the prognostic
424 algorithm, Radial Basis Function Support Vector Machine (RBF SVM) algorithm **(B)** NPV – negative
425 predictive value.

426

427

428

429

430 3.5. *In-vivo* validation of key diagnostic and prognostic biomarkers

431 We conducted a validation study of the panel of biomarkers to distinguish between different
432 stages of BE progression. Specifically, we performed RT-qPCR analysis to compare the
433 expression levels of these biomarkers in different cell lines: metaplasia (BAR-T and BAR-T10),
434 dysplasia (CP-B, CP-C and CP-D), and EAC (OE33, KYAE-1 and ESO26). Each biomarker was
435 tested with three technical replicates in each cell line.

436 For dysplasia diagnosis, we analyzed the expression of biomarkers in both metaplasia and
437 dysplasia cell lines (Figure S6). In evaluating EAC prognosis, we compared the expression levels
438 between metaplasia and EAC cell lines (Figure S7). Normalized expression values against
439 reference genes (*PGK1*, *ELF1* and *RPL13A*) highlighted significant differences in key markers.
440 For instance, biomarkers such as *IFI27* and *ATF3* differentiated metaplasia from dysplasia with
441 statistically significant p-values ($p = 0.009$ and $p = 0.003$, respectively), revealing their potential
442 utility in dysplasia diagnosis. Similarly, *CEBPB*, *SNAI1* and *CCN1* (alias *CYR61*) genes showed
443 significant expression changes between metaplasia and EAC (p-values of 0,031, 0.022 and
444 0.038, respectively). This supports their relevance for EAC prediction. The observed differential
445 expression patterns suggest that these biomarkers serve as valuable molecular tools for early
446 detection of dysplasia and the risk of progression to EAC, facilitating timely clinical intervention.
447 Interestingly, some of the top-performing genes, namely *IGHV3-43*, *IGHV4-31*, *IGHV3-53*, and
448 *PGC*, showed no detectable expression in cell lines. Since these genes ranked high according

449 to the diagnostic algorithm, we hypothesized that their expression originates from immune
450 cells, typically absent in cell lines. To investigate this further, we specifically tested these genes
451 in tissue samples from BE patients with and without dysplasia. Contrary to cell lines, the
452 expression of these genes was detectable in patient samples, supporting the notion that
453 immune cells- may play a critical role in BE progression (Figure S6).

454
455

456 **4. DISCUSSION:**

457 BE is the only known precursor to EAC, a malignancy with rising incidence and poor prognosis.
458 This underscores the need for more effective management methods, including assertive early
459 diagnosis of dysplasia and prognostic prediction within BE surveillance programs. While tools
460 incorporating biomarkers are continuously emerging, few have reached clinical validation and
461 implementation. Even fewer combine biomarkers with AI and those under clinical validation or
462 use, do not provide simultaneous detection of dysplasia and prognostic assessment. Moreover,
463 none can simultaneously achieve high sensitivity (recall) and high specificity.
464 In this study, we developed two algorithms to assist with the diagnosis of dysplasia, the
465 prognosis of BE, and ultimately the management of EAC. The genes of interest for dysplasia
466 detection (diagnostic algorithm) were newly identified from the raw data of three different
467 RNAseq datasets. Conversely, the algorithm developed for prognosis was based on a gene set
468 identified in a previous study [35]. For both applications, genes were ranked based on their F1
469 score, sensitivity (*aka* recall or true positive rate) and precision (*aka* positive predictive values)
470 in predicting conditions such as LGD BE and P-BE.
471 In high-risk disease detection cases such as dysplasia, recall is a more important evaluation
472 metric than precision because it can correctly identify all relevant positive cases (*i.e.*, samples
473 containing dysplasia or at high risk of progressing to EAC). However, precision, which is the
474 fraction of positive cases among all cases classified as positives by the model, is also crucial
475 because it emphasizes the correctness of positive predictions made by the model (*i.e.*,
476 measures how many cases are incorrectly classified as positive). In a situation where false
477 positives have significant implications, such as subjecting BE patients without dysplasia or with
478 a low risk of progression to unnecessary treatments or screening intervention, precision
479 matters. Since both high precision and high recall were desirable for the present study, the

480 ranking was based on the F1 score, which combines precision and recall using their harmonic
481 mean. Maximizing the F1 score implies maximizing both precision and recall simultaneously.
482 Performance metrics for each gene at its best threshold were high (see Supplementary Tables
483 5 and 6). However, specificity and NPV were higher for the prognostic genes, showing their
484 great potential to exclude patients who are not at risk for malignant progression.
485 To better explore the potential predictive value of the selected biomarkers, we trained
486 machine learning algorithms testing all possible combinations of biomarkers in each gene set.
487 The average metrics of the newly trained algorithms with combinations of biomarkers showed
488 increased predictive power (**Table 2**) compared to the predictive power of individual genes
489 (Supplementary Tables 5 and 6), which is expected in the context of complex gene interactions.
490 Finally, envisioning the clinical applicability of both algorithms, we evaluated the minimal
491 number of biomarkers necessary to maintain high-performance metrics (LSVM for diagnosis
492 and RBF SVM for prognosis) in each gene set. Both algorithms were tested with a decreasing
493 number of genes, and as depicted in **Figure 4**, a reduction in performance metrics was
494 observed when removing one gene from the diagnostic set and four genes from the
495 prognostic's gene set.
496 For diagnostic application, ten genes- *IGHV3-43*, *SLC38A4*, *PLL*, *CELA3A*, *IGHV4-31*, *TMPRSS5*,
497 *TP53*, *NR4A1*, *ATF3*, *IFI27*- were identified as the top candidates for dysplasia detection,
498 particularly for distinguishing between NDBE and LGD BE. These genes are associated with
499 different aspects of cancer biology, such as metabolism, cell invasion, and oncogenic
500 processes, suggesting their potential as biomarkers in the context of BE dysplasia [61–64].
501 Moreover, transcription factors such as *NR4A1* and *ATF3*, have been previously associated with
502 BE with LGD [65].

503 For the prognostic application, an RBF SVM algorithm was selected, which uses the expression
504 pattern of ten genes (*SNAI1*, *C1S*, *DUSP1*, *CEBPB*, *COL4A1*, *ZEB1*, *CEBPD*, *CCN1*, *LAMC1* and
505 *TWIST1*). Four of these genes – *SNAI1*, *COL4A1*, *ZEB1*, and *TWIST* – have been associated with
506 epithelial-to-mesenchymal transition [66]. *COL4A1*, *ZEB1*, and *TWIST1* have also been
507 described as potential screening biomarkers of BE malignant progression. *COL4A1* is
508 upregulated in EAC versus BE [67–69] and is associated with poor EAC prognosis [68], and it
509 predicts the response to immune checkpoint inhibitors in EAC [67]. Increased expression of
510 *ZEB1* has been associated with the repression of *CDH1* [70], which is associated with BE
511 progression to EAC [71–75]. *TWIST1* up-regulation was observed in at-risk BE samples years
512 before the emergence of any microscopic signs of malignancy (dysplasia/EAC) [35].

513 The genes *TP53* and *CDH1* were included in both gene sets to train the classifiers. *TP53* is
514 known for its role in BE malignant progression [76, 77], improved prediction of BE neoplastic
515 progression [78], increased risk of dysplasia when abnormally expressed, and improved intra-
516 observer agreement in dysplastic diagnosis [79]. *CDH1* has severely reduced or disorganized
517 expression during BE dysplastic progression [reviewed by [80] and an almost undetectable
518 expression in poorly differentiated EAC [71–75]. While *TP53* alone is insufficient for diagnostic
519 and for prognostic applications, it has been shown to have predictive value in combination with
520 other biomarkers in the diagnostic setting. These findings confirm the previously studied role
521 of *TP53* in the pathogenesis of BE dysplasia [81, 82]. Because *TP53* mutations are often
522 associated with a higher risk of progression in BE patients [83], further validation of this
523 biomarker at the molecular level, including its mutational status and RNA expression levels, is
524 warranted.

525 All metrics of both algorithms are higher when compared to currently available tools for risk
526 stratification, such as *TP53* immunohistochemistry (0.49 recall/sensitivity, 0.86 specificity [81])

527 and TissueCypher (0.55 recall/sensitivity, 0.82 specificity for high-intermediate risk class
528 55%/82%) [63]. Tools for dysplasia detection, such as Wats3D and Cytosponge-*TFF3*. are still
529 under prospective evaluation. Wats3D provides an incremental yield of 7% for any dysplasia
530 subtype but is negative for dysplasia in 62.5% of cases where an endoscopic biopsy
531 confirmation to compare with the gold standard revealed dysplasia [86]. The Cytosponge-*TFF3*
532 test when combined with a multidimensional biomarker panel and fitted into a regression
533 model was shown to be able to predict patients with dysplasia with good accuracy but further
534 validation is still needed [87]. Interestingly, in our top 45 genes for diagnostic application
535 (Supplementary Table 4), we have identified another trefoil factor, the *TFF2*, which is BE
536 related gene.

537 A preliminary *in vivo* validation of the selected diagnostic and prognostic biomarkers was
538 conducted by examining their expression in metaplasia, dysplasia and EAC-derived cell lines.
539 This validation confirmed their differential expression, highlighting their potential in
540 distinguishing BE progression stages. Exceptionally, *IGHV3-43*, *IGHV4-31*, *IGHV3-53*, and *PGC*
541 top-ranked genes were validated in FFPE samples from patients diagnosed with BE with and
542 without dysplasia due to their lack of expression in the cell lines. The absence of immune cells
543 in cell line cultures, which focus on epithelial cells, likely contributes to these findings. While
544 we cannot exclude that the used cell lines may exhibit genetic differences from the original
545 tissue, which potentially influences their molecular profiles [88], further clinical validation with
546 a selected cohort of patient samples is warranted and is currently underway.

547 No molecular tools are currently implemented in clinical practice for identifying LGD/HGD BE.
548 Dysplasia is a major biomarker in BE risk stratification, but it is often focal, making accurate
549 characterization of collected BE biopsy challenging [89], and leading to many cases of BE
550 classified as INDBE. INDBE is a management limbo for dysplasia, posing problems for clinicians.

551 Most clinical tools developed for BE focus on risk stratification (prognosis) [28, 37–39, 84, 85,
552 90] and have a high specificity (identify and correctly exclude BE patients not at risk of
553 progression). Simultaneously, these tools have a lower recall/sensitivity indicating their
554 performance drops in detecting BE patients at true risk of progression.

555 New tests that aim for high recall and sensitivity are vital to avoid missing unacceptable true
556 positive cases of LGD or HGD, as well as patients at risk of progression. However, these tests
557 must also maintain high precision and high sensitivity to avoid incorrectly including patients
558 not having dysplasia or having a low risk of progression. This balance can improve surveillance
559 of high-risk patients while reducing unnecessary procedures for low-risk patients, ultimately
560 lowering patient management costs. Our approach, which combines machine learning
561 algorithms with gene expression signatures, represents a promising breakthrough in
562 healthcare. It has the potential to significantly enhance both the diagnosis and prognosis of
563 dysplasia by delivering high recall and precision into clinical practice.

564

565 5. CONCLUSIONS:

566 This study not only identified biomarkers and developed algorithms to detect LGD in BE
567 biopsies and predict the progression of BE to EAC, but also paved the way for creating new *in-*
568 *vitro* laboratory tests for the diagnosis and prognosis of BE. Both algorithms were developed
569 using datasets from public databases analyzing tissue samples obtained during routine
570 endoscopy.

571 For the prediction of BE malignant progression, an LSVM algorithm featuring the identification
572 of LGD was trained while an RBF SVM algorithm was trained for the prediction of BE malignant
573 progression. Both algorithms reached high-performance metrics. To our knowledge, no
574 existing tools can simultaneously detect dysplasia and assess the risk of progression with such
575 high precision and recall.

576 Validation of the biomarkers and algorithms presented in this study in an independent test and
577 validation patient cohort is currently under consideration. Additionally, while no other known
578 risk factors (epidemiologic, clinical, histologic) have been combined with the presented
579 biomarkers, incorporating patient demographic and clinical information could further enhance
580 the predictive value of the gene expression algorithms. Future algorithm developments will
581 address this issue, demonstrating how such combinations can significantly boost their
582 predictive power.

583

584

585

586

587

588

589 **FUNDING INFORMATION:** This work was supported by Ophiomics – Precision Medicine own
590 funding.

591

592 **AUTHOR CONTRIBUTIONS:** **Migla Miskinyte:** Conceptualization, Formal analysis,
593 Investigation, Writing – review & editing. **Benilde Pondevca:** Conducting experiments, Writing
594 – review & editing. **José B. Pereira-Leal:** Conceptualization, review & editing. **Joana Cardoso:**
595 Conceptualization, Validation, Investigation, Writing – review & editing. All authors have read
596 and agreed to the published version of the manuscript.

597

598 **CONFLICT OF INTEREST STATEMENT:** The work described here is subject to European Patent
599 Application No. 24172031.7; JPL, JC declare an ownership interest in the company Ophiomics
600 – Precision Medicine. MM, BP are employees at Ophiomics – Precision Medicine.

601

Supplementary Material:

Supplementary Table 1 – List of genes used for differential expression analysis across multiple datasets

Supplementary Table 2 – List of differentially expressed genes between NDBE and LGDBE groups, E_MTAB_4054 dataset.

Supplementary Table 3 – List of differentially expressed genes between NDBE and LGDBE groups, GSE58963dataset.

Supplementary Table 4 – Top 45 candidate genes.

Supplementary Table 5 – Top 15 genes for prognostic: individual genes' predictive metrics.

N	Gene	Threshold	F1-score	Recall	Precision	Specificity	NPV	Accuracy	FPR	TP	FP	TN	FN
1	<i>COL4A1</i>	9.34	0.84	0.75	0.94	0.98	0.89	0.91	0.02	49	3	132	16
2	<i>LAMC1</i>	9.19	0.84	0.88	0.81	0.90	0.94	0.90	0.10	57	13	122	8
3	<i>CEBPB</i>	9.75	0.84	0.80	0.88	0.95	0.91	0.90	0.05	52	7	128	3
4	<i>CCN1</i>	7.68	0.84	0.88	0.80	0.90	0.94	0.89	0.10	57	14	121	8
5	<i>SNAI1</i>	6.61	0.82	0.97	0.71	0.81	0.98	0.86	0.19	63	26	109	2
6	<i>C1S</i>	9.65	0.83	0.83	0.83	0.92	0.92	0.89	0.08	54	11	124	11
7	<i>ZEB1</i>	7.66	0.80	0.91	0.72	0.83	0.95	0.86	0.17	59	23	112	6
8	<i>CEBPD</i>	9.32	0.81	0.83	0.79	0.90	0.92	0.88	0.10	54	14	121	11
9	<i>DUSP1</i>	10.46	0.81	0.82	0.80	0.90	0.91	0.88	0.10	53	13	122	12
10	<i>VWF</i>	8.88	0.77	0.80	0.74	0.87	0.90	0.85	0.13	52	18	117	13
11	<i>TWIST</i>	5.15	0.67	1.00	0.51	0.53	1.00	0.69	0.47	65	63	72	0
12	<i>PLPP3</i>	8.68	0.68	0.94	0.53	0.60	0.95	0.71	0.40	61	54	81	4
13	<i>ACTN1</i>	9.82	0.73	0.74	0.72	0.86	0.87	0.82	0.14	48	19	116	17
14	<i>CDH1</i>	5.70	0.50	1.00	0.33			0.33	1.00	65	135	0	0
15	<i>TP53</i>	7.12	0.62	0.89	0.48	0.53	0.91	0.65	0.47	58	64	71	7

Supplementary Table 6 – Top 18 genes for diagnostic: individual genes' predictive metrics.

N	Gene	Threshold	F1-score	Recall	Precision	Specificity	NPV	Accuracy	FPR	TP	FP	TN	FN
1	<i>SLC38A4</i>	-0.21	0.84	0.90	0.78	0.52	0.73	0.77	0.48	36	10	11	4
2	<i>TMPRSS5</i>	-1.14	0.82	0.98	0.71	0.24	0.83	0.72	0.76	39	16	5	1
3	<i>EGR3</i>	0.04	0.82	0.98	0.71	0.24	0.83	0.72	0.76	39	16	5	1
4	<i>TP53</i>	5.05	0.82	0.95	0.72	0.29	0.75	0.72	0.71	38	15	6	2
5	<i>FOSB</i>	1.15	0.82	1.00	0.69	0.14	1.00	0.70	0.86	40	18	3	0
6	<i>NR4A1</i>	3.22	0.81	1.00	0.68	0.10	1.00	0.69	0.90	40	19	2	0
7	<i>SFTPB</i>	-0.10	0.80	0.95	0.69	0.19	0.67	0.69	0.81	38	17	4	2
8	<i>IFI27</i>	3.87	0.79	0.98	0.66	0.05	0.50	0.66	0.95	39	20	1	1
9	<i>PLLP</i>	5.66	0.78	0.95	0.66	0.05	0.33	0.64	0.95	38	20	1	2
10	<i>CELA3A</i>	-3.84	0.78	0.95	0.66	0.05	0.33	0.64	0.95	38	20	1	2
11	<i>ATF3</i>	2.94	0.78	0.95	0.66	0.05	0.33	0.64	0.95	38	20	1	2
12	<i>IGHV3-43</i>	-1.38	0.78	0.95	0.66	0.05	0.33	0.64	0.95	38	20	1	2
13	<i>CDH1</i>	8.61	0.78	0.95	0.66	0.05	0.33	0.64	0.95	38	20	1	2
14	<i>PGC</i>	2.48	0.74	0.88	0.64	0.05	0.17	0.59	0.95	35	20	1	5
15	<i>GKN2</i>	-0.11	0.74	0.88	0.64	0.05	0.17	0.59	0.95	35	20	1	5
16	<i>IGHV4-31</i>	-0.91	0.74	0.88	0.64	0.05	0.17	0.59	0.95	35	20	1	5
17	<i>IGHV3-53</i>	-0.92	0.74	0.88	0.64	0.05	0.17	0.59	0.95	35	20	1	5
18	<i>PNLIPRP1</i>	-2.73	0.72	0.83	0.63	0.10	0.22	0.57	0.90	33	19	2	7

Supplementary Table 7 – Gene frequency in the best-performing algorithms for diagnostic (LSVM = 1115)

Gene	Count	Frequency (%)
<i>IGHV3-43</i>	1115	100.00
<i>SLC38A4</i>	1097	98.39
<i>PLLP</i>	1072	96.14
<i>CELA3A</i>	923	82.78
<i>IGHV4-31</i>	850	76.23
<i>TMPRSS5</i>	646	57.94
<i>TP53</i>	645	57.85
<i>NR4A1</i>	633	56.77
<i>ATF3</i>	625	56.05
<i>IFI27</i>	581	52.11
<i>PGC</i>	530	47.53
<i>GKN2</i>	490	43.95
<i>PNLIPRP1</i>	470	42.15
<i>SFTPB</i>	460	41.26
<i>CDH1</i>	379	34.00
<i>FOSB</i>	369	33.10
<i>EGR3</i>	355	31.84
<i>IGV3-53</i>	330	29.60

Supplementary Table 8 – Gene frequency in the best-performing algorithms for prognostic (RBF SVM = 5794)

Gene	Count	Frequency (%)
<i>SNAI1</i>	5545	95.70245
<i>DUSP1</i>	3837	66.22368
<i>CEBPB</i>	3791	65.42975
<i>C1S</i>	3672	63.37591
<i>COL4A1</i>	3409	58.83673
<i>LAMC1</i>	3157	54.4874
<i>CEBPD</i>	3121	53.86607
<i>ZEB1</i>	3105	53.58992
<i>TWIST1</i>	2940	50.74215
<i>CCN1</i>	2913	50.27615
<i>TP53</i>	2872	49.56852
<i>ACTN1</i>	2609	45.02934
<i>VWF</i>	2592	44.73593
<i>PLPP3</i>	2336	40.31757
<i>CDH1</i>	1140	19.67553

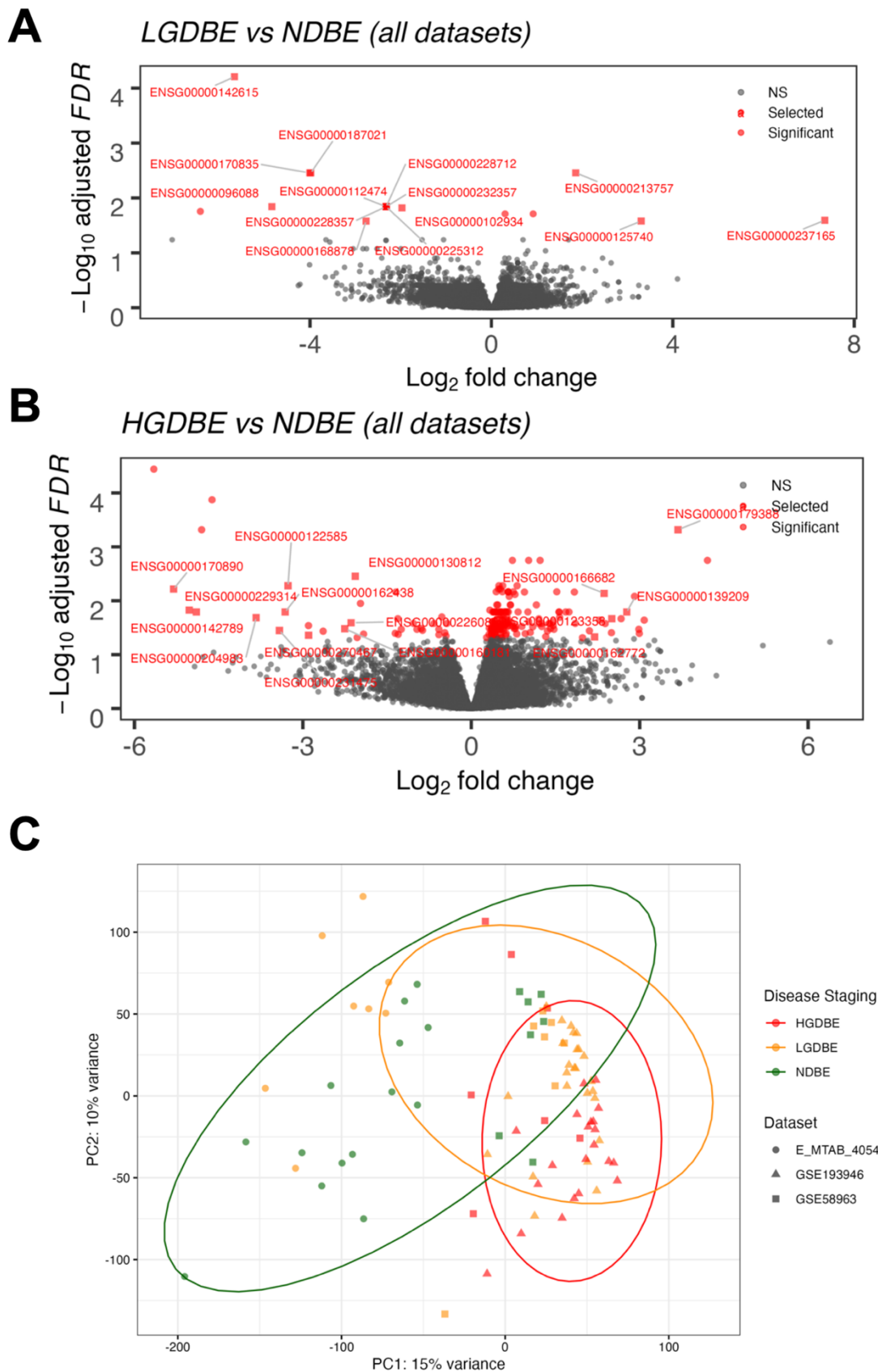


Figure S1. Identification of 30 biomarkers using all datasets.

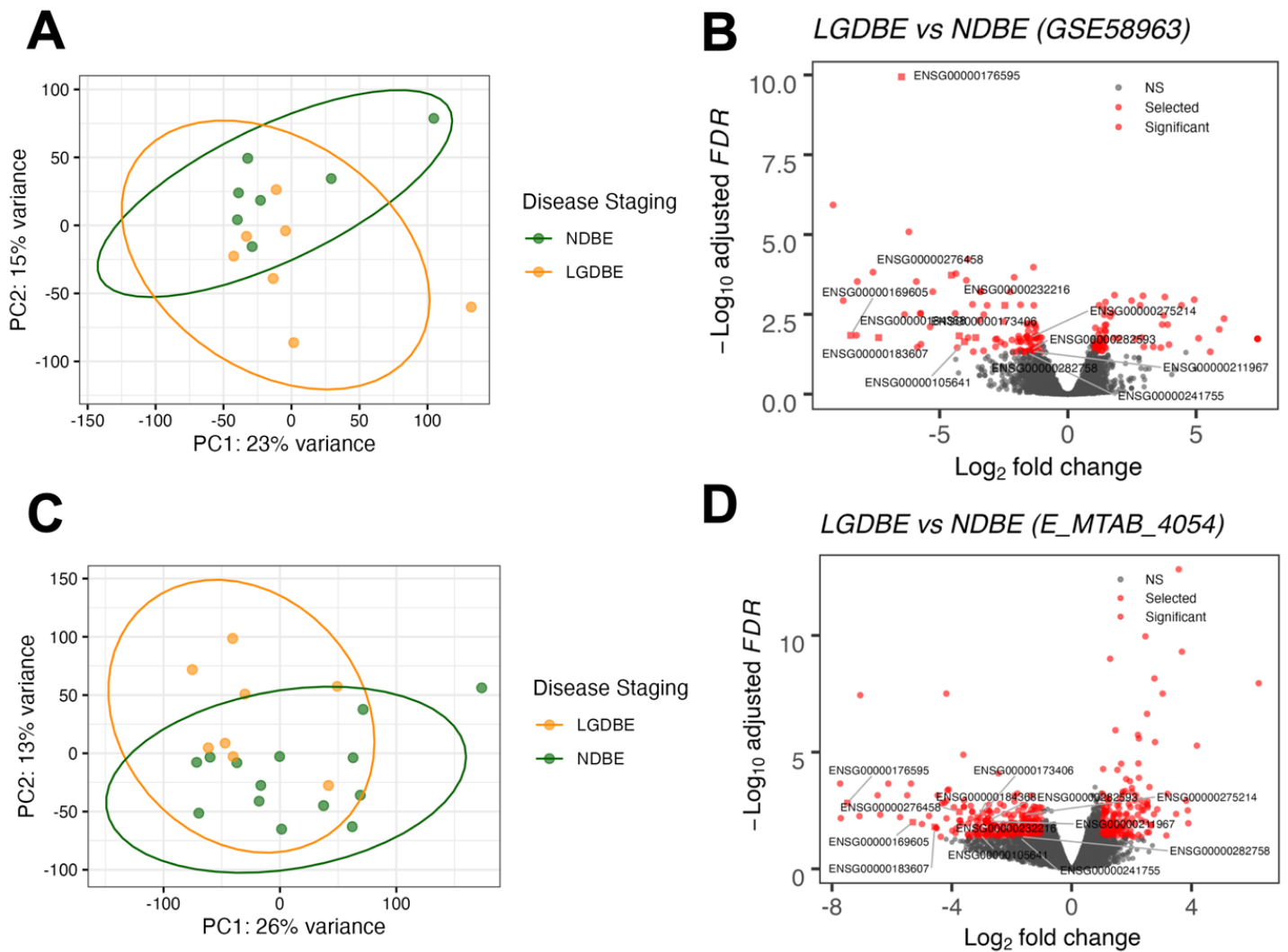


Figure S2. Identification of additional 13 biomarkers by analyzing datasets separately.

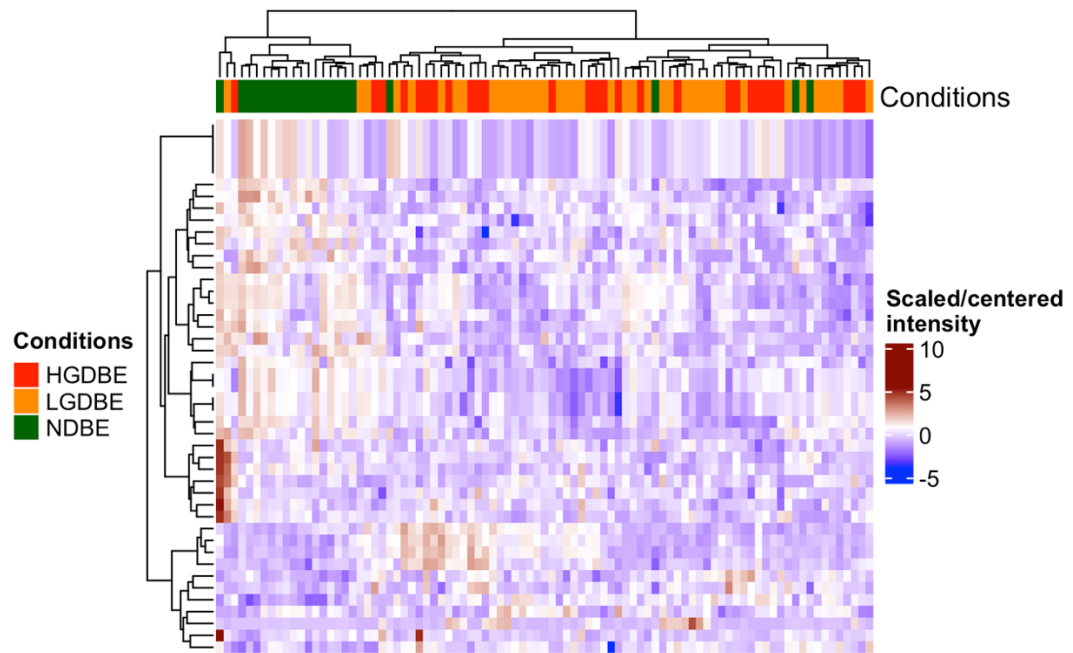


Figure S3. Heatmap of top 45 genes selected for diagnostics.

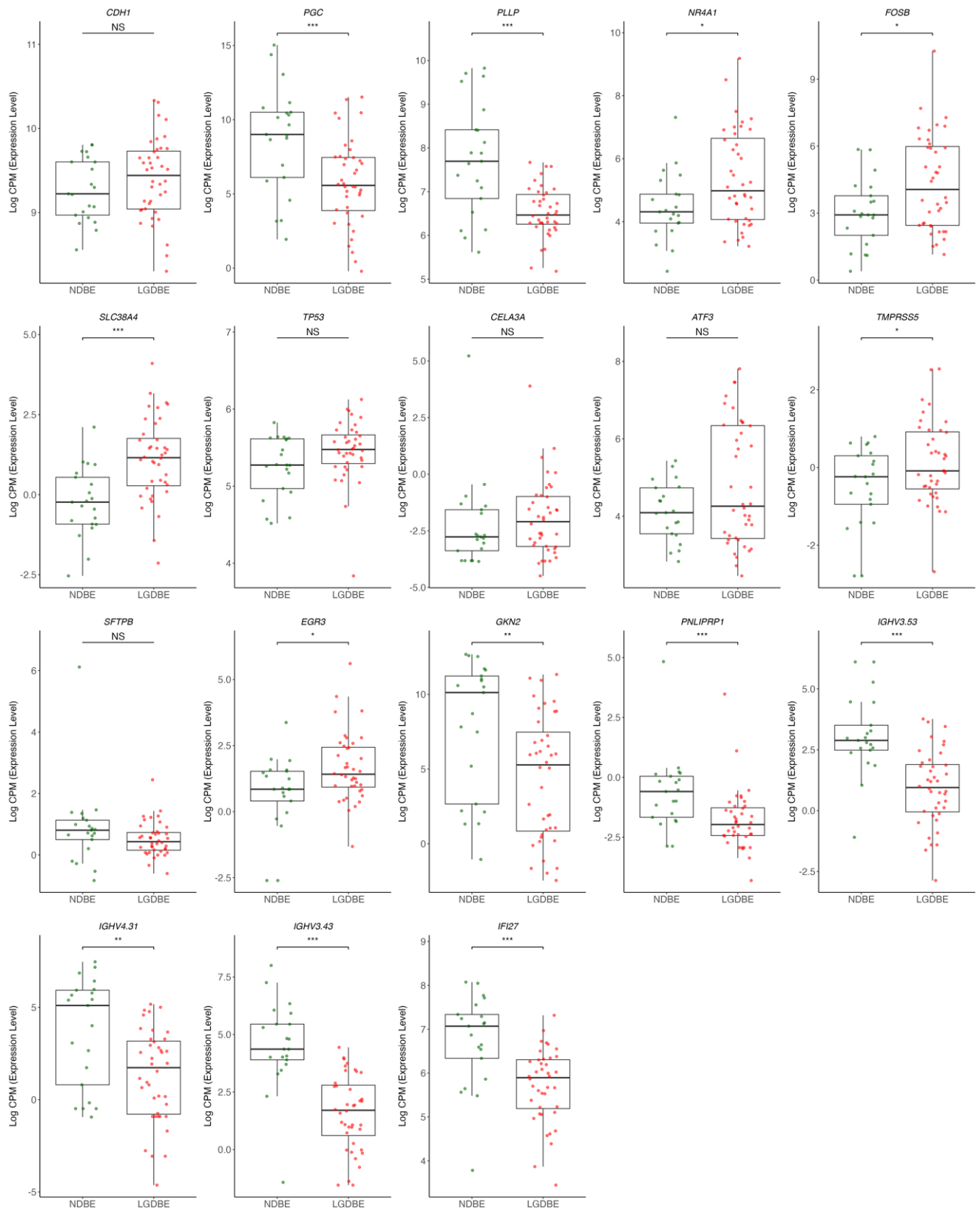


Figure S4. Boxplots of gene expression levels of genes potentially associated with low-grade dysplasia. Comparison of each gene expression in NDBE and LGDBE samples obtained from a total of 3 datasets. *** adj. $p < 0.001$; ** adj. $p < 0.01$; * adj. $p < 0.05$; NS adj. $p > 0.05$.

All rights reserved. No reuse allowed without permission.

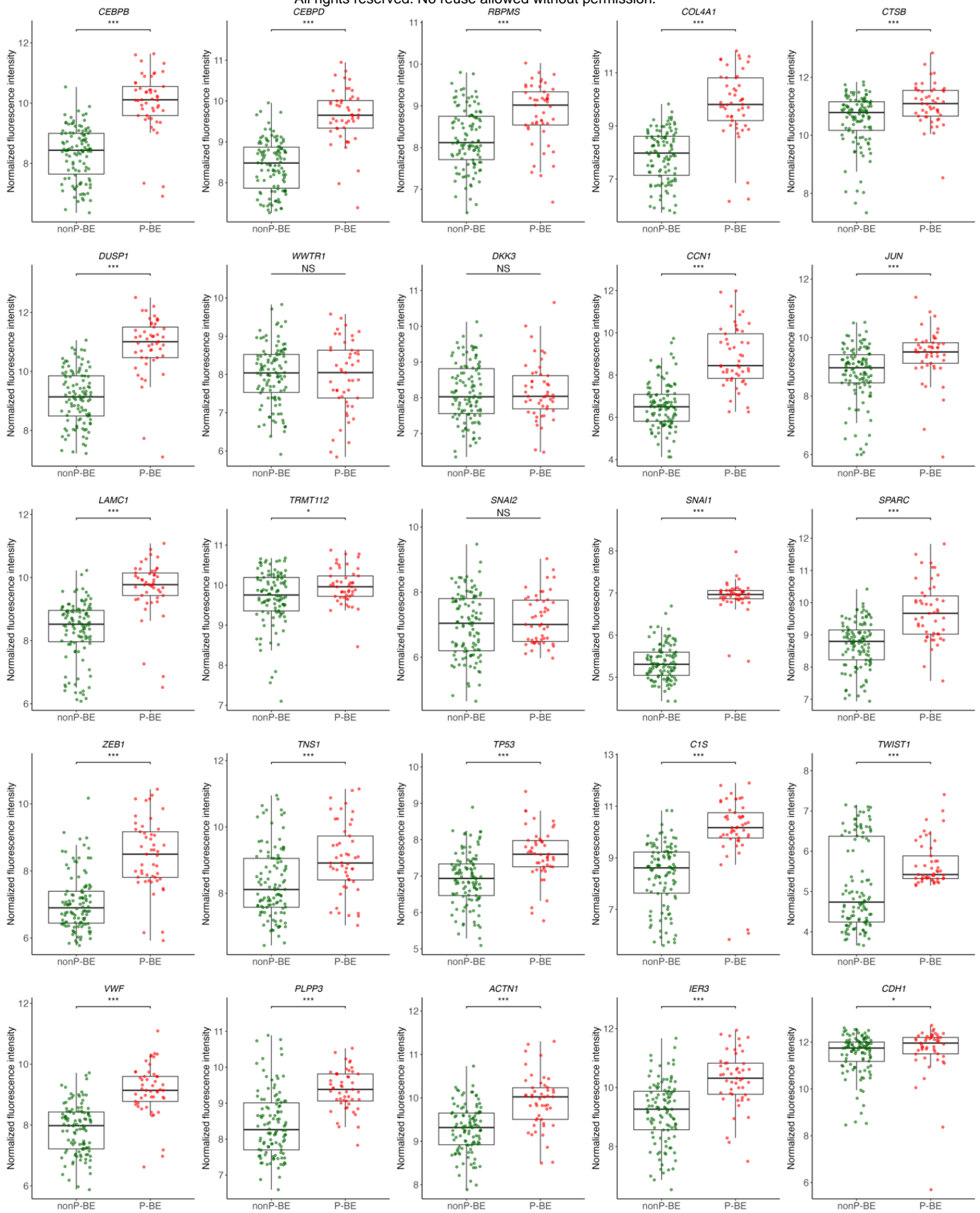


Figure S5. Boxplots of gene expression levels of genes potentially associated with prognosis. Comparison of each gene expression in nonP-BE and P-BE samples obtained from a total of 9 datasets. *** adj. $p < 0.001$; ** adj. $p < 0.01$; * adj. $p < 0.05$; NS adj. $p > 0.05$.

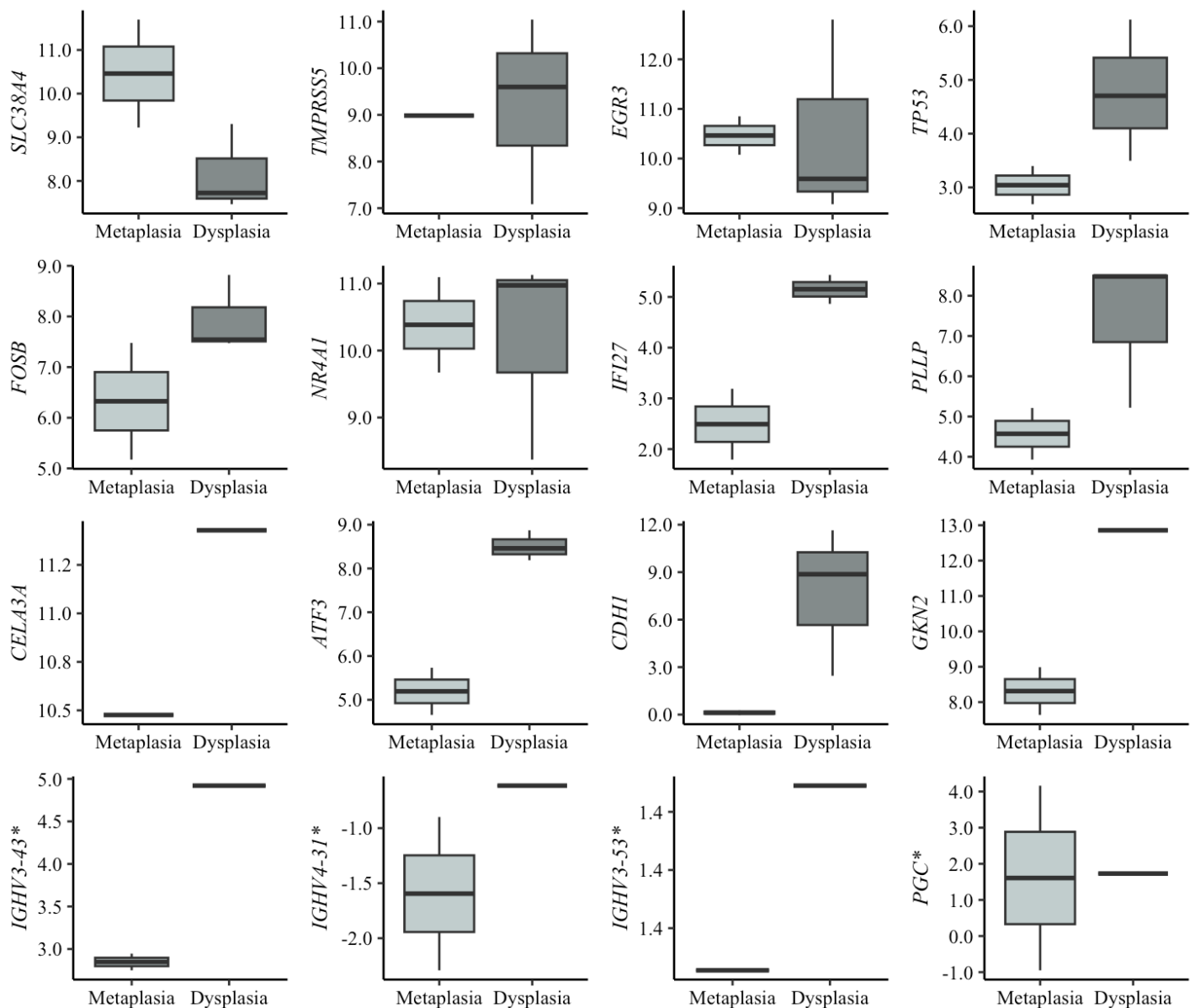


Figure S6. Boxplots of gene expression levels of diagnostic genes. Comparison of gene expression levels between cell lines representing metaplasia and dysplasia, with expression levels normalized to reference genes (*PGK1*, *ELF1*, and *RPL13A*). * Genes were additionally tested in FFPE samples from patients diagnosed with BE with and without dysplasia.

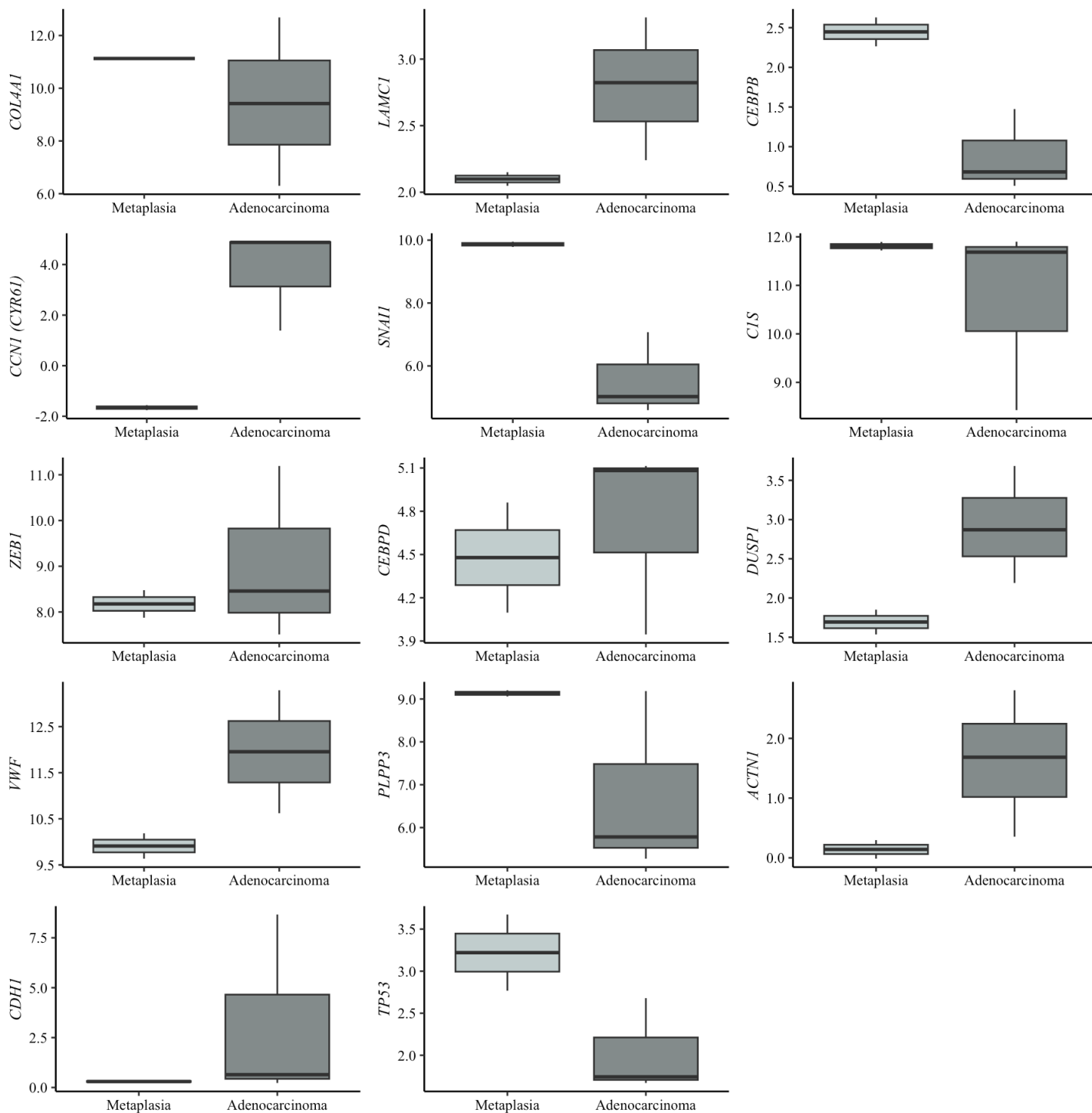


Figure S7. Boxplots of gene expression levels of prognostic genes. Comparison of gene expression levels between cell lines representing metaplasia and EAC, with expression levels normalized to reference genes (*PGK1*, *ELF1*, and *RPL13A*).

REFERENCES:

1. Spechler SJ, Souza RF. Barrett's Esophagus. *New England Journal of Medicine*. 2014;371:836–45.
2. Choi KKH, Sanagapalli S. Barrett's esophagus: Review of natural history and comparative efficacy of endoscopic and surgical therapies. *World J Gastrointest Oncol*. 2022;14:568–86.
3. Klavan H, Russell MB, Macklin J, Lee E, Aslanian HR, Muniraj T. Barrett's esophagus: A comprehensive review for the internist. *Disease-a-Month*. 2018;64:471–87.
4. Killcoyne S, Fitzgerald RC. Evolution and progression of Barrett's oesophagus to oesophageal cancer. *Nature Reviews Cancer*. 2021;21:731–41.
5. Malfertheiner P, Nocon M, Vieth M, Stolte M, Jaspersen D, Koelz HR, et al. Evolution of gastro-oesophageal reflux disease over 5 years under routine medical care--the ProGERD study. *Aliment Pharmacol Ther*. 2012;35:154–64.
6. Fabian T, Leung A. Epidemiology of Barrett's Esophagus and Esophageal Carcinoma. *Surgical Clinics of North America*. 2021;101:381–9.
7. Cook MB, Shaheen NJ, Anderson LA, Giffen C, Chow WH, Vaughan TL, et al. Cigarette smoking increases risk of Barrett's esophagus: An analysis of the Barrett's and esophageal adenocarcinoma consortium. *Gastroenterology*. 2012;142:744–53.
8. Sinha DN, Abdulkader RS, Gupta PC. Smokeless tobacco-associated cancers: A systematic review and meta-analysis of Indian studies. *Int J Cancer*. 2016;138:1368–79.
9. Kamat P, Wen S, Morris J, Anandasabapathy S. Exploring the Association Between Elevated Body Mass Index and Barrett's Esophagus: A Systematic Review and Meta-Analysis. *Annals of Thoracic Surgery*. 2009;87:655–62.
10. Andrici J, Tio M, Cox MR, Eslick GD. Hiatal hernia and the risk of Barrett's esophagus. *Journal of Gastroenterology and Hepatology (Australia)*. 2013;28:415–31.
11. Mittal SK, Abdo J, Adrien MP, Bayu BA, Kline JR, Sullivan MM, et al. Current state of prognostication, therapy and prospective innovations for Barrett's-related esophageal adenocarcinoma: a literature review. *J Gastrointest Oncol*. 2021;12:1197–214.
12. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71:209–49.
13. Morgan E, Soerjomataram I, Runggay H, Coleman HG, Thrift AP, Vignat J, et al. The Global Landscape of Esophageal Squamous Cell Carcinoma and Esophageal Adenocarcinoma Incidence and Mortality in 2020 and Projections to 2040: New Estimates From GLOBOCAN 2020. *Gastroenterology*. 2022;163:649-658.e2.
14. Then EO, Lopez M, Saleem S, Gayam V, Sunkara T, Culliford A, et al. Esophageal cancer: An updated surveillance epidemiology and end results database analysis. *World J Oncol*. 2020;11:55–64.
15. Fabian T, Leung A. Epidemiology of Barrett's Esophagus and Esophageal Carcinoma. *Surgical Clinics of North America*. 2021;101:381–9.
16. Hamade N, Vennelaganti S, Parasa S, Vennalaganti P, Gaddam S, Spaander MCW, et al. Lower Annual Rate of Progression of Short-Segment vs Long-Segment Barrett's Esophagus to Esophageal Adenocarcinoma. *Clin Gastroenterol Hepatol*. 2019;17:864.
17. Li N, Yang WL, Cai MH, Chen X, Zhao R, Li MT, et al. Burden of gastroesophageal reflux disease in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of disease study 2019. *BMC Public Health*. 2023;23:1–13.

18. Zhang D, Liu S, Li Z, Wang R. Global, regional and national burden of gastroesophageal reflux disease, 1990–2019: update from the GBD 2019 study. *Ann Med*. 2022;54:1372–84.
19. Spechler SJ, Sharma P, Souza RF, Inadomi JM, Shaheen NJ. American gastroenterological association technical review on the management of Barrett’s esophagus. *Gastroenterology*. 2011;140.
20. Lee SW, Lien HC, Chang C Sen, Lin MX, Chang CH, Ko CW. Benefits of the Seattle biopsy protocol in the diagnosis of Barrett’s esophagus in a Chinese population. *World J Clin Cases*. 2018;6:753–8.
21. Eluri S, Shaheen NJ. Barrett’s Esophagus: Diagnosis and Management. *Gastrointest Endosc*. 2017;85:889.
22. Visrodia K, Singh S, Krishnamoorthi R, Ahlquist DA, Wang KK, Iyer PG, et al. Magnitude of Missed Esophageal Adenocarcinoma After Barrett’s Esophagus Diagnosis: A Systematic Review and Meta-analysis. *Gastroenterology*. 2016;150:599-607.e7.
23. Sharma P. Low-grade dysplasia in Barrett’s esophagus. *Gastroenterology*. 2004;127:1233–8.
24. Runge TM, Abrams JA, Shaheen NJ. Epidemiology of Barrett’s Esophagus and Esophageal Adenocarcinoma. *Gastroenterology Clinics of North America*. 2015;44:203–31.
25. Alshelleh M, Inamdar S, McKinley M, Stewart M, Novak JS, Greenberg RE, et al. Incremental yield of dysplasia detection in Barrett’s esophagus using volumetric laser endomicroscopy with and without laser marking compared with a standardized random biopsy protocol. *Gastrointest Endosc*. 2018;88:35–42.
26. Thota PN, Kistangari G, Esnakula AK, Gonzalo DH, Liu X-L. Clinical significance and management of Barrett’s esophagus with epithelial changes indefinite for dysplasia. *World J Gastrointest Pharmacol Ther*. 2016;7:406–11.
27. Shaheen NJ, Falk GW, Iyer PG, Souza RF, Yadlapati RH, Sauer BG, et al. Diagnosis and Management of Barrett’s Esophagus: An Updated ACG Guideline. *Am J Gastroenterol*. 2022;117:559–87.
28. Vaughan TL, Onstad L, Dai JY. Interactive decision support for esophageal adenocarcinoma screening and surveillance. *BMC Gastroenterol*. 2019;19.
29. Galipeau PC, Li X, Blount PL, Maley CC, Sanchez CA, Odze RD, et al. NSAIDs Modulate CDKN2A, TP53, and DNA Content Risk for Progression to Esophageal Adenocarcinoma. *PLoS Med*. 2007;4:e67.
30. Trindade AJ, McKinley MJ, Alshelleh M, Levi G, Stewart M, Quinn KJ, et al. Mutational load may predict risk of progression in patients with Barrett’s oesophagus and indefinite for dysplasia: A pilot study. *BMJ Open Gastroenterol*. 2019;6.
31. Maley CC, Galipeau PC, Li X, Sanchez CA, Paulson TG, Blount PL, et al. The Combination of Genetic Instability and Clonal Expansion Predicts Progression to Esophageal Adenocarcinoma. 2004.
32. Mokrowiecka A, Wierzchniewska-Ławska A, Smolarz B, Romanowicz-Makowska H, Małecka-Panas E. P16 gene mutations in Barrett’s esophagus in gastric metaplasia - Intestinal metaplasia - Dysplasia - Adenocarcinoma sequence. *Adv Med Sci*. 2012;57:71–6.
33. Paulson TG, Maley CC, Li X, Li H, Sanchez CA, Chao DL, et al. Chromosomal instability and copy number alterations in Barrett’s esophagus and esophageal adenocarcinoma. *Clinical Cancer Research*. 2009;15:3305–14.
34. Merlo LMF, Shah NA, Li X, Blount PL, Vaughan TL, Reid BJ, et al. A comprehensive survey of clonal diversity measures in Barrett’s esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer Prevention Research*. 2010;3:1388–97.

35. Cardoso J, Mesquita M, Dias Pereira A, Bettencourt-Dias M, Chaves P, Pereira-Leal JB. CYR61 and TAZ upregulation and focal epithelial to mesenchymal transition may be early predictors of Barrett's esophagus malignant progression. *PLoS One*. 2016;11.
36. Selaru FM, Zou T, Xu Y, Shustova V, Yin J, Mori Y, et al. Global gene expression profiling in Barrett's esophagus and esophageal cancer: a comparative analysis using cDNA microarrays. *Oncogene*. 2022;21:475–8.
37. Moinova HR, Laframboise T, Lutterbaugh JD, Chandar AK, Dumot J, Faulx A, et al. Identifying DNA methylation biomarkers for non-endoscopic detection of Barrett's esophagus. 2018.
38. Jin Z, Cheng Y, Gu W, Zheng Y, Sato F, Mori Y, et al. A multicenter, double-blinded validation study of methylation biomarkers for progression prediction in Barrett's esophagus. *Cancer Res*. 2009;69:4112–5.
39. Abdo J, Wichman CS, Dietz NE, Ciborowski P, Fleegel J, Mittal SK, et al. Discovery of novel and clinically relevant markers in formalin-fixed paraffin-embedded esophageal cancer specimen. *Front Oncol*. 2018;8 MAY.
40. Tan JL, Chinnaratha MA, Woodman R, Martin R, Chen HT, Carneiro G, et al. Diagnostic Accuracy of Artificial Intelligence (AI) to Detect Early Neoplasia in Barrett's Esophagus: A Non-comparative Systematic Review and Meta-Analysis. *Front Med (Lausanne)*. 2022;9:890720.
41. Fouad YM, Mostafa I, Yehia R, EL-Khayat H. Biomarkers of Barrett's esophagus. *World J Gastrointest Pathophysiol*. 2014;5:450.
42. Honing J, Fitzgerald RC. Categorizing Risks within Barrett's Esophagus To Guide Surveillance and Interception; Suggesting a New Framework. *Cancer Prev Res (Phila)*. 2023;16:313–20.
43. Mejza M, Małecka-Wojcieszko E. Diagnosis and Management of Barrett's Esophagus. *J Clin Med*. 2023;12.
44. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26:139–40.
45. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28:882.
46. Watts GS, Tran NL, Berens ME, Bhattacharyya AK, Nelson MA, Montgomery EA, et al. Identification of Fn14/TWEAK receptor as a potential therapeutic target in esophageal adenocarcinoma. *Int J Cancer*. 2007;121:2132–9.
47. Jabeen A, Ahmad N, Raza K. Machine Learning-based state-of-the-art methods for the classification of RNA-Seq data. In: *Classification in Bioapps: Automation of Decision Making*. Springer; 2018. p. 133–72.
48. Pirooznia M, Yang JY, Qu MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*. 2008;9 SUPPL. 1.
49. Peixoto C, Lopes MB, Martins M, Casimiro S, Sobral D, Grosso AR, et al. Identification of biomarkers predictive of metastasis development in early-stage colorectal cancer using network-based regularization. *BMC Bioinformatics*. 2023;24.
50. MacCarthy FP, Duggan SP, Feighery R, O J, Ravi N, Kelleher D, et al. 708 IL-1B and SERPINA-3 Are Novel Markers of Aggressive Barrett's Oesophagus Phenotype Using RNA Deep Sequencing Analysis. *Gastroenterology*. 2014;146:S-122.
51. Maag JLV, Fisher OM, Levert-Mignon A, Kaczorowski DC, Thomas ML, Hussey DJ, et al. Novel aberrations uncovered in Barrett's esophagus and esophageal adenocarcinoma using whole transcriptome sequencing. *Molecular Cancer Research*. 2017;15:1558–69.

52. Zhang Y, Weh KM, Howard CL, Riethoven JJ, Clarke JL, Lagisetty KH, et al. Characterizing isoform switching events in esophageal adenocarcinoma. *Mol Ther Nucleic Acids*. 2022;29:749–68.
53. Kimchi ET, Posner MC, Park JO, Darga TE, Kocherginsky M, Karrison T, et al. Progression of Barrett's Metaplasia to Adenocarcinoma Is Associated with the Suppression of the Transcriptional Programs of Epidermal Differentiation.
54. Ostrowski J, Mikula M, Karczmariski J, Rubel T, Wyrwicz LS, Bragoszewski P, et al. Molecular defense mechanisms of Barrett's metaplasia estimated by an integrative genomics. *J Mol Med*. 2007;85:733–43.
55. Stairs DB, Nakagawa H, Klein-Szanto A, Mitchell SD, Silberg DG, Tobias JW, et al. Cdx1 and c-Myc foster the initiation of transdifferentiation of the normal esophageal squamous epithelium toward Barrett's esophagus. *PLoS One*. 2008;3.
56. Silvers AL, Lin L, Bass AJ, Chen G, Wang Z, Thomas DG, et al. Decreased selenium-binding protein 1 in esophageal adenocarcinoma results from posttranscriptional and epigenetic regulation and affects chemosensitivity. *Clinical Cancer Research*. 2010;16:2009–21.
57. Di Pietro M, Lao-Sirieix P, Boyle S, Cassidy A, Castillo D, Saadi A, et al. Evidence for a functional role of epigenetically regulated midcluster HOXB genes in the development of Barrett esophagus. *Proc Natl Acad Sci U S A*. 2012;109:9077–82.
58. Wang Q, Ma C, Kemmner W. Wdr66 is a novel marker for risk stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma. *BMC Cancer*. 2013;13.
59. Hyland PL, Hu N, Rotunno M, Su H, Wang C, Wang L, et al. Global changes in gene expression of Barrett's esophagus compared to normal squamous esophagus and gastric cardia tissues. *PLoS One*. 2014;9.
60. Cummings LC, Thota PN, Willis JE, Chen Y, Cooper GS, Furey N, et al. A nonrandomized trial of vitamin D supplementation for Barrett's esophagus. *PLoS One*. 2017;12.
61. Kastelein F, Biermann K, Steyerberg EW, Verheij J, Kalisvaart M, Looijenga LHJ, et al. Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with Barrett's oesophagus. *Gut*. 2013;62:1676–83.
62. Li J, Li M han, Wang T tian, Liu X ning, Zhu X ting, Dai Y zhang, et al. SLC38A4 functions as a tumour suppressor in hepatocellular carcinoma through modulating Wnt/ β -catenin/MYC/HMGCS2 axis. *Br J Cancer*. 2021;125:865–76.
63. Shulgin AA, Lebedev TD, Prassolov VS, Spirin P V. Plasmolipin and Its Role in Cell Processes. *Mol Biol*. 2021;55:773.
64. Jadhav K, Zhang Y. Activating transcription factor 3 in immune response and metabolic regulation. *Liver Res*. 2017;1:96–102.
65. Maag JLV, Fisher OM, Levert-Mignon A, Kaczorowski DC, Thomas ML, Hussey DJ, et al. Novel aberrations uncovered in Barrett's esophagus and esophageal adenocarcinoma using whole transcriptome sequencing. *Molecular Cancer Research*. 2017;15:1558–69.
66. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol*. 2014;15:178–96.
67. Li K, Duan P, He H, Du R, Wang Q, Gong P, et al. Construction of the Interaction Network of Hub Genes in the Progression of Barrett's Esophagus to Esophageal Adenocarcinoma. *J Inflamm Res*. 2023;16:1533–51.
68. Qi W, Li R, Li L, Li S, Zhang H, Tian H. Identification of key genes associated with esophageal adenocarcinoma based on bioinformatics analysis. *Ann Transl Med*. 2021;9:1711–1711.

69. Nancarrow DJ, Clouston AD, Smithers BM, Gotley DC, Drew PA, Watson DI, et al. Whole genome expression array profiling highlights differences in mucosal defense genes in Barrett's esophagus and esophageal adenocarcinoma. *PLoS One*. 2011;6.
70. Zhang Q, Agoston AT, Pham TH, Zhang W, Zhang X, Huo X, et al. Acidic Bile Salts Induce Epithelial to Mesenchymal Transition via VEGF Signaling in Non-Neoplastic Barrett's Cells. *Gastroenterology*. 2019;156:130-144.e10.
71. Yao C, Li Y, Luo L, Xiong Q, Zhong X, Xie F, et al. Identification of miRNAs and genes for predicting Barrett's esophagus progressing to esophageal adenocarcinoma using miRNA-mRNA integrated analysis. *PLoS One*. 2021;16.
72. Kalatskaya I. Overview of major molecular alterations during progression from Barrett's esophagus to esophageal adenocarcinoma. *Ann N Y Acad Sci*. 2016;1381:74–91.
73. Darlavoix T, Seelentag W, Yan P, Bachmann A, Bosman FT. Altered expression of CD44 and DKK1 in the progression of Barrett's esophagus to esophageal adenocarcinoma. *Virchows Archiv*. 2009;454:629–37.
74. Feith M, Stein HJ, Mueller J, Siewert JR. Malignant degeneration of Barrett's esophagus: the role of the Ki-67 proliferation fraction, expression of E-cadherin and p53. 2004.
75. Falkenback D, Nilbert M, Öberg S, Johansson J. Prognostic value of cell adhesion in esophageal adenocarcinomas. *Diseases of the Esophagus*. 2008;21:97–102.
76. Paulson TG, Galipeau PC, Oman KM, Sanchez CA, Kuhner MK, Smith LP, et al. Somatic whole genome dynamics of precancer in Barrett's esophagus reveals features associated with disease progression. *Nat Commun*. 2022;13.
77. Pinto R, Hauge T, Jeanmougin M, Pharo HD, Kresse SH, Honne H, et al. Targeted genetic and epigenetic profiling of esophageal adenocarcinomas and non-dysplastic Barrett's esophagus. *Clin Epigenetics*. 2022;14.
78. Redston M, Noffsinger A, Kim A, Akarca FG, Rara M, Stapleton D, et al. Abnormal TP53 Predicts Risk of Progression in Patients With Barrett's Esophagus Regardless of a Diagnosis of Dysplasia. *Gastroenterology*. 2022;162:468–81.
79. Januszewicz W, Pilonis ND, Sawas T, Phillips R, O'Donovan M, Miremadi A, et al. The utility of P53 immunohistochemistry in the diagnosis of Barrett's oesophagus with indefinite for dysplasia. *Histopathology*. 2022;80:1081–90.
80. Kalatskaya I. Overview of major molecular alterations during progression from Barrett's esophagus to esophageal adenocarcinoma. *Ann N Y Acad Sci*. 2016;1381:74–91.
81. Kastelein F, Biermann K, Steyerberg EW, Verheij J, Kalisvaart M, Looijenga LHJ, et al. Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with Barrett's oesophagus. *Gut*. 2013;62:1676–83.
82. Li S, Hoefnagel SJM, Krishnadath KK. Molecular Biology and Clinical Management of Esophageal Adenocarcinoma. *Cancers* 2023, Vol 15, Page 5410. 2023;15:5410.
83. Redston M, Noffsinger A, Kim A, Akarca FG, Rara M, Stapleton D, et al. Abnormal TP53 Predicts Risk of Progression in Patients With Barrett's Esophagus Regardless of a Diagnosis of Dysplasia. *Gastroenterology*. 2022;162:468.
84. Eluri S, Brugge WR, Daglilar ES, Jackson SA, Styn MA, Callenberg KM, et al. The presence of genetic mutations at key loci predicts progression to esophageal adenocarcinoma in Barrett's esophagus. *American Journal of Gastroenterology*. 2015;110:828–34.
85. Iyer PG, Codipilly DC, Chandar AK, Agarwal S, Wang KK, Leggett CL, et al. Prediction of Progression in Barrett's Esophagus Using a Tissue Systems Pathology Test: A Pooled Analysis of International Multicenter Studies. *Clinical Gastroenterology and Hepatology*. 2022. <https://doi.org/10.1016/j.cgh.2022.02.033>.

86. Codipilly DC, Krishna Chandar A, Wang KK, Katzka DA, Goldblum JR, Thota PN, et al. Wide-area transepithelial sampling for dysplasia detection in Barrett's esophagus: a systematic review and meta-analysis. *Gastrointest Endosc.* 2022;95:51-59.e7.
87. Ross-Innes CS, Chettouh H, Achilleos A, Galeano-Dalmau N, Debiram-Beecham I, MacRae S, et al. Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. *Lancet Gastroenterol Hepatol.* 2017;2:23–31.
88. Panda A, Bhanot G, Ganesan S, Bajpai M. Gene expression in barrett's esophagus cell lines resemble esophageal squamous cell carcinoma instead of esophageal adenocarcinoma. *Cancers (Basel).* 2021;13:5971.
89. Odze RD. Update on the Diagnosis and Treatment of Barrett Esophagus and Related Neoplastic Precursor Lesions. *Arch Pathol Lab Med.* 2007;132:1577–85.
90. Kaul V, Gross S, Corbett FS, Malik Z, Smith MS, Tofani C, et al. Clinical utility of wide-area transepithelial sampling with three-dimensional computer-assisted analysis (WATS3D) in identifying Barrett's esophagus and associated neoplasia. *Diseases of the Esophagus.* 2020;33.